LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

“To propose a new algorithm for improving software architecture with clustering algorithm”

A Dissertation Report

Submitted by

Shilpa Sharma

11301074

To

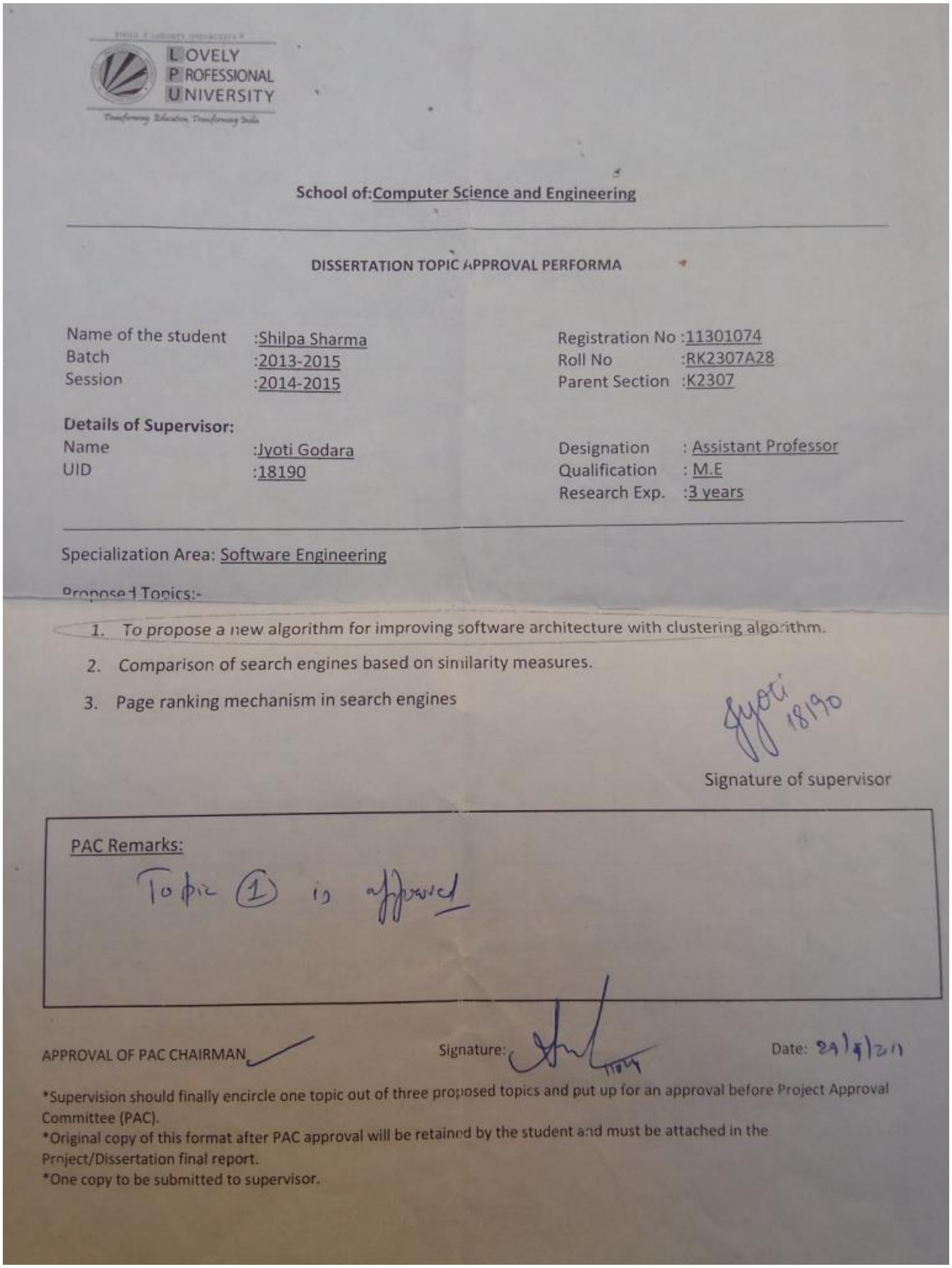
Department of Computer Science and Engineering

In partial fulfilment of the Requirement for the
Award of the Degree of

Master of Technology In
Computer Science and Engineering

Under the guidance of
Ms. JYOTI GODARA
Asst. Professor, LPU.
(MAY 2015)

PAC APPROVAL FORM



LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: Computer Science and Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the student :Shilpa Sharma

Batch :2013-2015

Session :2014-2015

Registration No :11301074

Roll No :RK2307A28

Parent Section :K2307

Details of Supervisor:

Name :Jyoti Godara

UID :18190

Designation : Assistant Professor

Qualification : M.E

Research Exp. :3 years

Specialization Area: Software Engineering

Proposed Topics:-

1. To propose a new algorithm for improving software architecture with clustering algorithm.
2. Comparison of search engines based on similarity measures.
3. Page ranking mechanism in search engines

Jyoti 18190

Signature of supervisor

PAC Remarks:

Topic ① is approved

APPROVAL OF PAC CHAIRMAN ✓

Signature:

Date: 24/4/2014

*Supervision should finally encircle one topic out of three proposed topics and put up for an approval before Project Approval Committee (PAC).

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to supervisor.

ABSTRACT

The size and difficulty of business potential software systems are continually rising. This means that the task of supervision of a large software project is becoming even more demanding, particularly in light of high turnover of skilled human resources. Software clustering approaches can help with the task of understanding large, complicated software systems by automatically decomposing them into smaller, easier-to-manage subsystems. In this paper we recognize significant study instructions in the area of software clustering that require further attention in order to develop more effective and efficient clustering methodologies for software engineering. To that end, we first there the state of the art in software clustering research. We consider the methods of clustering that have received the most attention from the research community and outline their strengths and weaknesses. Our paper defines every stages of a clustering algorithm separately. We also present the most important approaches for evaluating the value of software clustering.

CERTIFICATE

This is to certify that **Shilpa Sharma** has completed M.Tech Dissertation proposal titled "**To propose a new algorithm for improving software architecture with clustering algorithm"** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engineering.

Date:

Signature of advisor

Name: **Jyoti Godara**

ACKNOWLEDGEMENT

First and foremost, I want to thank the Department of CSE of Lovely Professional University for giving me permission to begin Thesis in first instance, to do necessary research work and to use required data. I would like to acknowledge the assistance provided to me by the library staff of L.P.U.

Inspiration to action is the most important ingredient required throughout the task. I am deeply indebted to my mentor Ms. Jyoti Godara whose help, stimulating suggestions and encouragement helped me in all the time of research.

I express my gratitude to my parents for being a continuous source of encouragement and for their financial aids given to me. Finally, I would like to express my gratitude to all those who helped and supported me.

Shilpa Sharama

DECLARATION

I hereby declare that the dissertation entitled, **to propose new algorithm form improving software architecture with clustering algorithm”** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Investigator
Regd.no: 11301074

TABLE OF CONTENTS

LIST OF FIGURES

Chapter-1

INTRODUCTION

1.1 Software engineering-Software engineering is the branch of study and application of engineering to the software which initially design development of the same software by using various technique or updations, maintenance of the software. Software engineering deals with the all kind of software production, design to coding, software accuracy and deals with the complexity of any software system. Software industry is moving very fast in the current scenario. Even big industries spend large some of amount on their software engineer for the software development. [1]

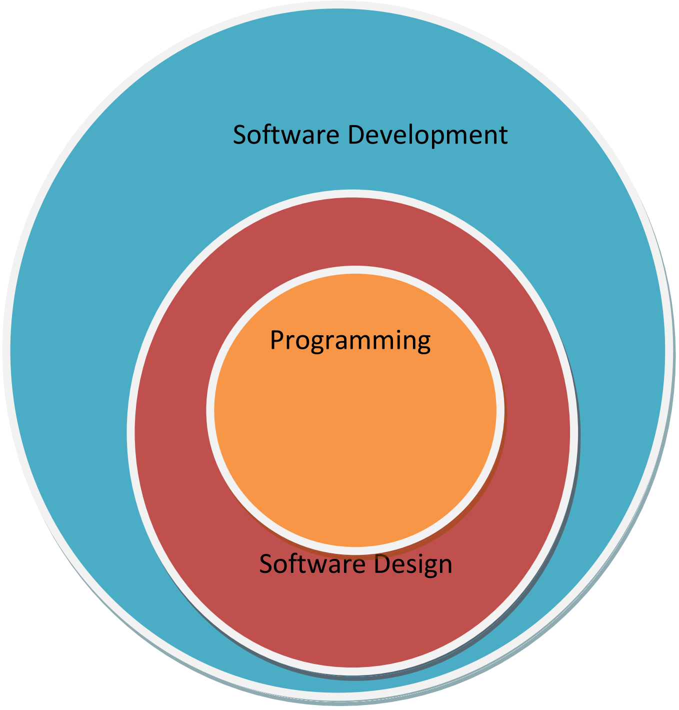


Fig 1.1. Software Engineering Module

Basically software Engineering is a study and integral part of engineering which include design, development and maintenance of the system. Software Engineers are applied principles to the software engineering to develop, design, test and maintain any software.

1.2 Software development life cycle: - software development life cycle is a methodology which introduces the different phases of a software product. In SDLC we usually passes from the various phases that is project definition, then requirement of user, requirement of system, analysis and design, implementation and sustainment. In SDLC we need to check the requirement of user and to design the same by fulfilling requirement of user. Later on we just coding the program then testing of the system is one of the important phase in SDLC. Here you can able to find any bug or error in the earlier stage, so that you can fix it. After completion the testing of software it is get ready to implement.

The different SDLC phases are:

1. Analysis

2. Design

3. Implementation

4. Testing

5. Maintenance

First of all it, feasibility study will be done at the initial state. After that information will be collected and gathered according to the requirement. IT means in this phase requirement specification and analysis will be done. SRS document all are design for design phase, Next phase is design phase in which DFD, flow chart are designed as a model of specification. Toy model, prototype, spiral model all are made in this phase.

After that Implementation part is done. In this DFD and flowcharts are converted into a programming language. Any platform can be used for programming part.

Testing is the next phase in which different types of testing are done to check error and bugs in the system Maintenance phase is the last phase to maintain the software with regression testing.

Figure refer to next page represent the diagrammatical presentation of all involved stages in Growth of particular software.

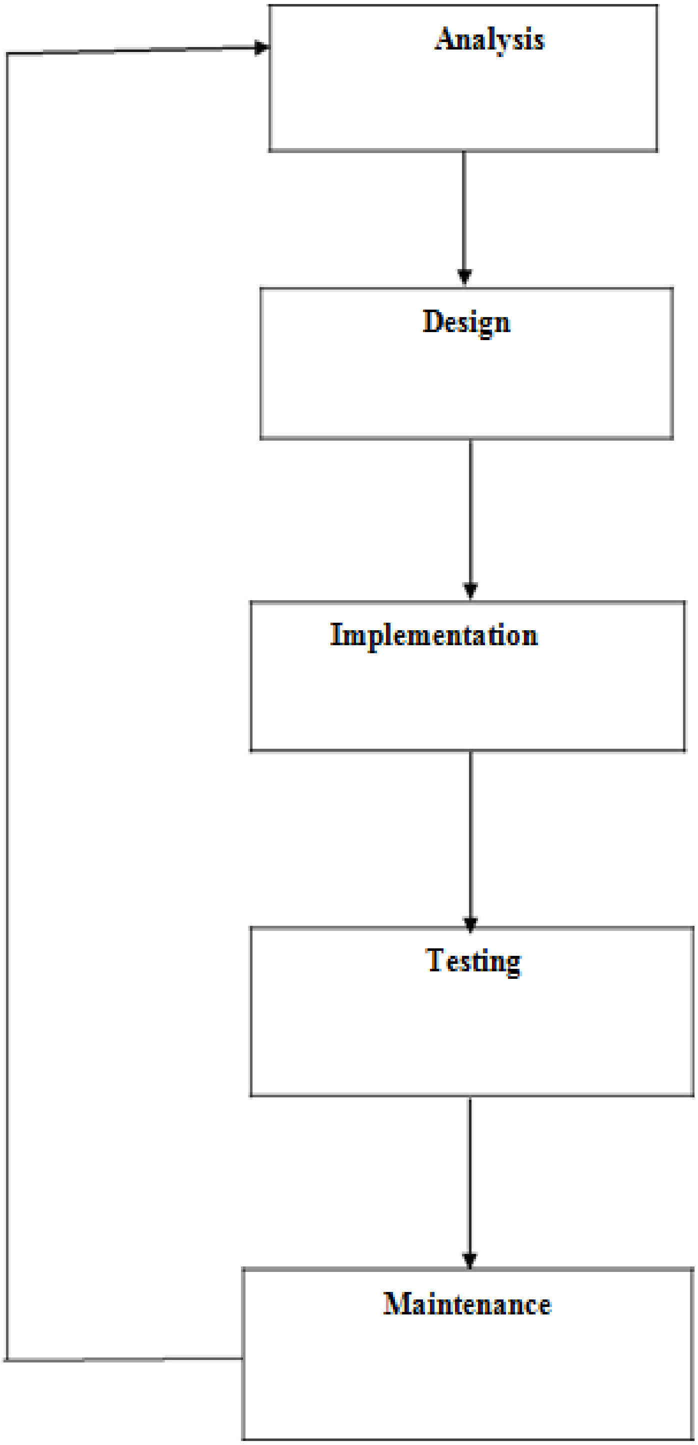


Fig 1.2 the phase of SDLC

1.3 Architecture of Software:

1.3.1 Architecture

Architecture is professional who design software or other computer application, or who prepare plans and strategy who upgrade it from time to time or according or user requirement. In other words software architecture is a process to defining a proposed solution that fulfils the entire technical and operational requirement offered to the user.

1.3.2 Software Component Models and Technologies

We discussed above, a software model identify and set standards for composition of and interaction among different software element. Such technique, various tools of software and new up gradation is often implemented. Component based software engineering is a dedicated software product that supports the use of particular software component model. The component model may be defined as a process where standard for component implementation, Documentation and distribution at each level of software system. It contain various tool like Run time environment for component execution and development of component, composition then deployment.

Examples of the Component model are Corba component model, EJB (environment java beans) and COM+ model.

COM+ is an expansion of COM incorporating hold for services, such as transactional processing and communication queuing, that are usually used in spread information systems. These services are not programmatically invoked from inner side of the components. Rather, declarative attributes can be combining with components and applications, specifying which services can or must be provided and at which level. The COM+ run-time system uses this information to cut off component connections and insert system calls as essential. This allows existing COM components to be clearly increased with, for instance, transactional processing and used as part of COM+ applications.One more model given that parallel services is Sun’s Enterprise JavaBeans (EJB) which is based on Java but not on the abovementioned JavaBeans model. The required service levels 7 for a set of EJB components are uttered declaratively in a file called a deployment descriptor.

After deployment, each of the objects implemented by the components, generally called beans, live inside an EJB container, which also contains objects generated from the deployment descriptor. Clients invoke a bean’s operations via these generated objects, which make sure the right service levels. Unlike JavaBeans, beans in EJB do not communicate via events.

There are two major types’ beans. Entity beans are used to summarize access to database records. An entity bean may apply its own persistence management or let the container manage persistence as specified by the deployment descriptor. Session beans, which may be state full or stateless, represent dealings sessions with clients. Message-driven beans can be seen as a particular kind of stateless session beans that symbolize asynchronous interaction session. A session bean may organize transactions or leave that to the container. EJB requires the Java 2 Enterprise Edition (J2EE) platform.

1.4 Clustering

The basic thought behind clustering is to create the groups together for analogous object or for similar purposes. In other word clustering may b define as a portioning of data contain into subset or in the small size cluster. Clustering wide use of algorithm (K-Mean, C-Mean etc.)

Clustering is procedure in which similar data elements are combined together and dissimilar are removed, in clustering different documents are grouped in a single groups. In this same documents or say similar documents are grouped in a same cluster. Many advantages are there but alternative advantage of clustering is that document will not misplace. In case a document is misplaced then it can be easily found by using clustering algorithms.

Unsupervised classification includes clustering. Sorting refers to a technique that allocates data objects to a set of classes. Unsupervised is defined as clustering does not rely on the already defined classes [21].

Unsupervised learning is that in which learning is by observations. It diverse from model reformation in the region of information called differentiates examination and choice examination which categorize the objects from a given set of object.

It analyzed of cluster mostly the traditional subjects in the data mining arena. It is considered as the very first stage in information detection. Clustering is defined as the exercise of assembly data items into a set of distinct classes, called clusters .Now substances inside a class have high similarity to each other in the interim objects in distinct classes are more complementary .[22]

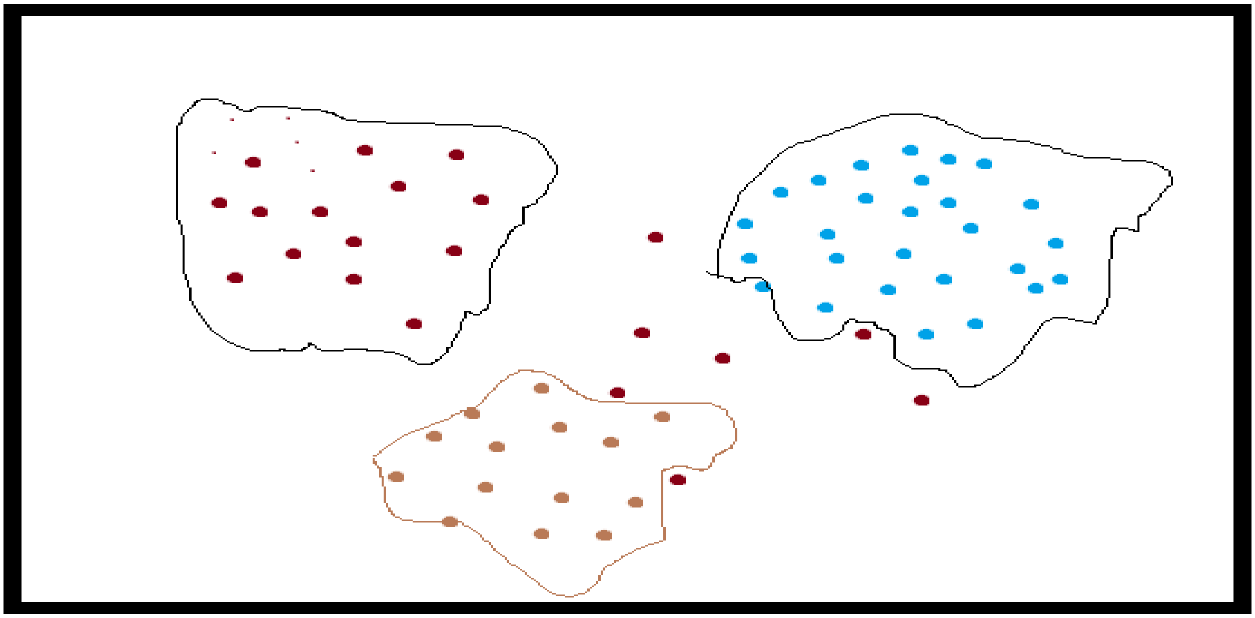


Fig.1.3 clustering of similar functions [23]

There are many clustering algorithms used for clustering. The major fundamental clustering methods can be classified into following categories [23]:

1. Partitioning Methods:-The general criterion for partitioning could be a mixture of high sameness of the samples within clusters with high difference between distinct clusters. Most strategies of the partitioning square measure distance-based. Given k, the amount of partitions to construct, a partitioning methodology creates associate degree initial partitioning and so uses associate degree unvaried reassign technique that tries to enhance the partitioning by moving objects from one cluster to a different. during a sensible partitioning the objects within the same cluster square measure shut or associated with one another whereas objects totally different in several in numerous} clusters square measure way apart or different. Most applications adopt well-liked heuristic strategies like greedy approaches just like the k-means and k-medoids algorithms that increasingly improve the bunch quality and approach a neighbourhood optimum. These bunch strategies works well for locating spherical –shaped clusters in little to medium size databases. During this construct a partition of a knowledge set containing n objects into a group of k clusters, thus to reduce a criterion .The goal is, given a k, notice a partition of k clusters that optimizes the chosen partitioning criterion. Here k could be a input parameters. E.g. K-mean and K-centroid [20].

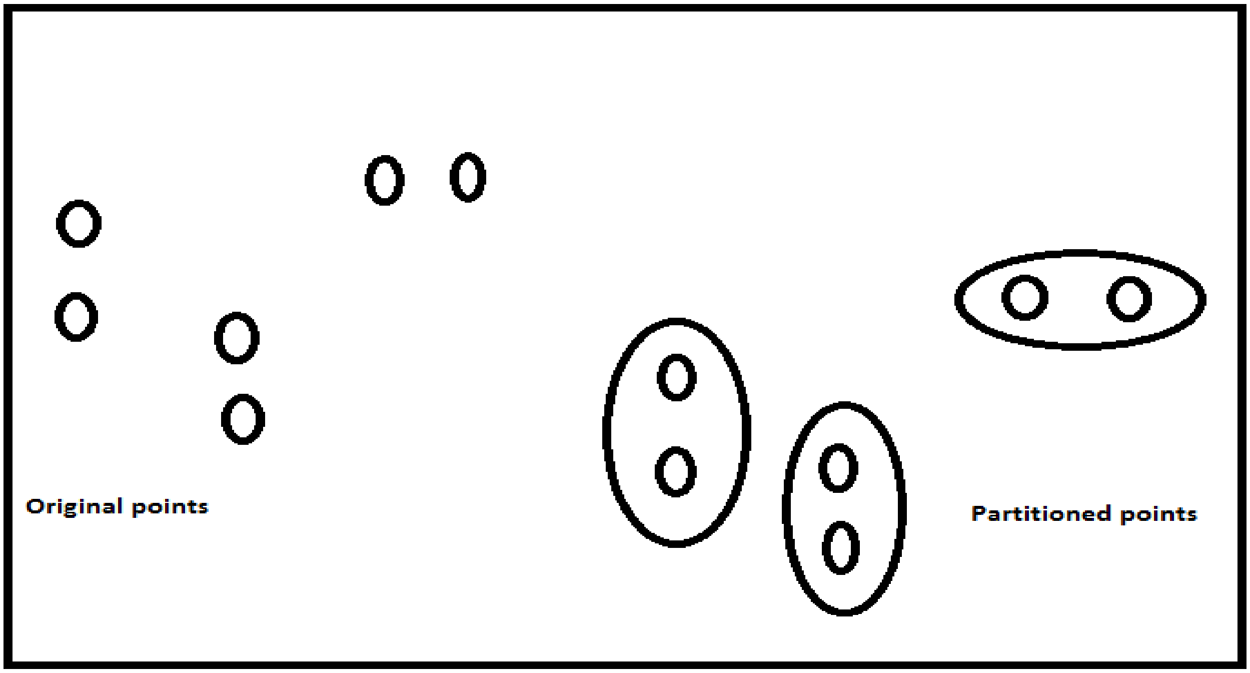


Fig.1.4 Partitioning Clustering [20]

2. Hierarchical Methods: during this methodology hierarchical decomposition of the given set of knowledge objects is made. It will be classified as being either agglomerated or dissentious supported however hierarchical decomposition is created. agglomerated approach is that the bottom up approach starts with every object forming a separate cluster.

Hierarchical algorithms produce a hierarchical decomposition of the given knowledge set of knowledge objects. The hierarchical decomposition is depicted by a tree structure, referred to as dendrogram. It doesn't would like clusters as inputs. During this style of bunch it's potential to look at partitions at completely different level of granularities victimization differing types of K. E.g. Flat bunch [2].

It then merges teams near each other till all the teams area unit unified into one. Dissentious approach is prime down approach starts with all the clusters within the same cluster then in every iteration step a cluster is split into smaller clusters till every object is object in one cluster.

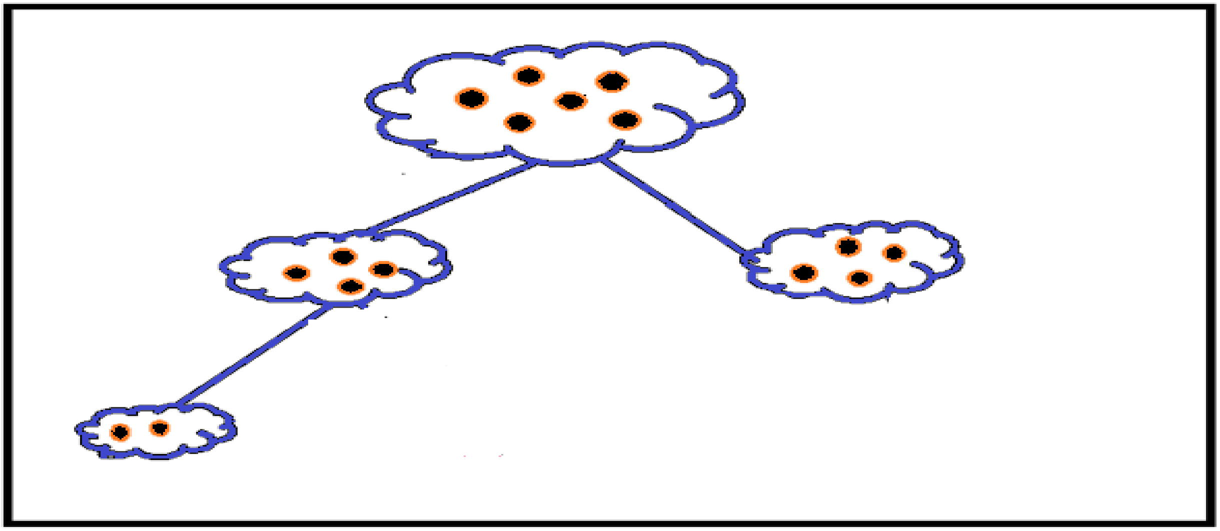


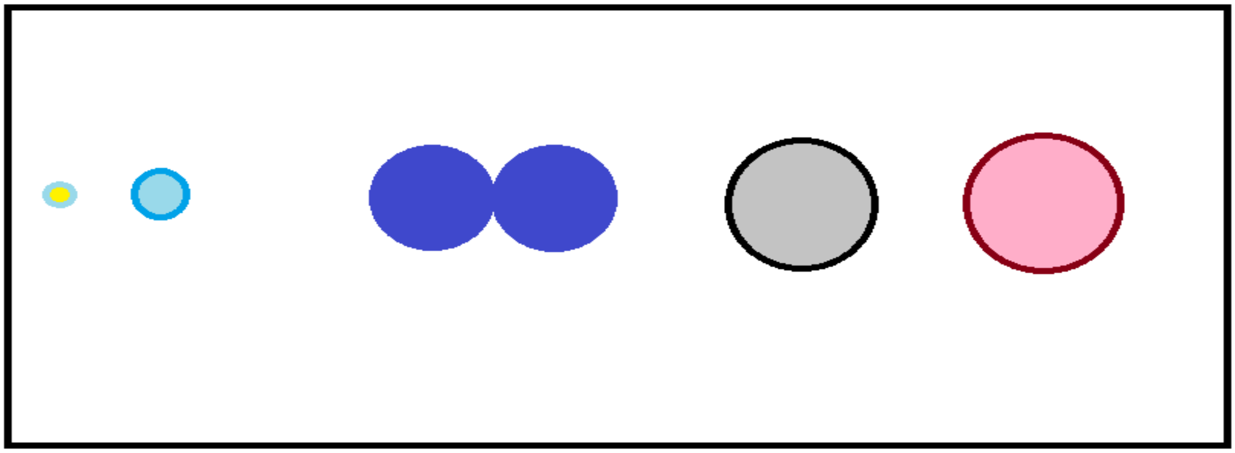
Fig.1.5 Hierarchal Clustering [2]

3. Density Based Methods:- Most partitioning methods cluster objects based on distance between objects. It is one time scan technique. According to the regions which grow with high density t finds clusters. Clusters are high density area than remaining data set. Density is the number of objects in a cluster. It finds arbitrary shaped clusters. It is applicable to spatial data. This technique is of two types-

1. Based on density function

2. Based upon connectivity points

It can handle noise. Moreover it doesn't require any specified number of cluster. But it fails to work upon high density of data.



Fig, 1.6 Density based clustering [20]

4. **Well shaped Cluster:** A cluster is a package of nodes in which any node in a cluster is closer or more similar to every other node in the same cluster than to any node not in the cluster. Sometimes threshold can be used to specify similarity or closeness between the nodes in cluster [19].

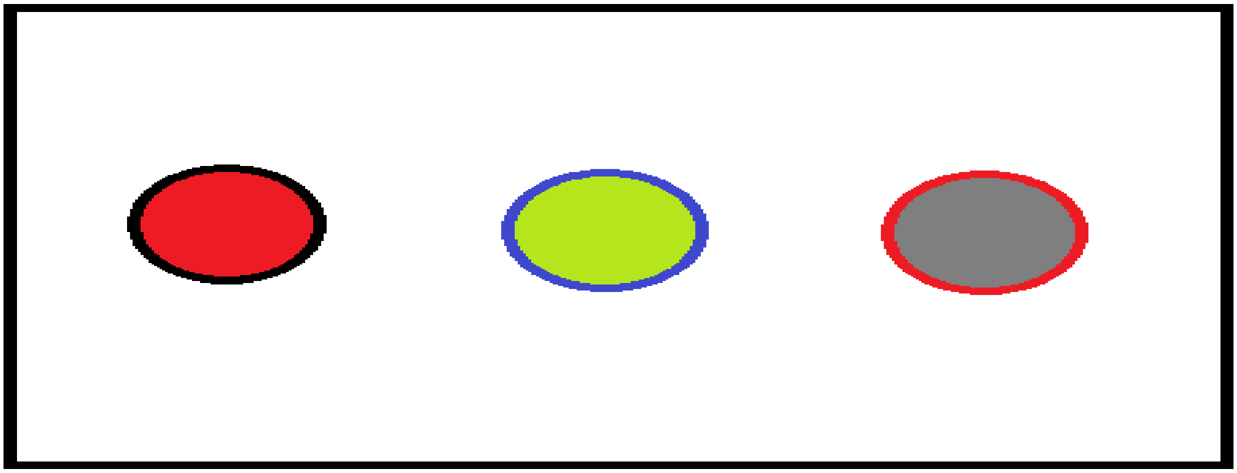
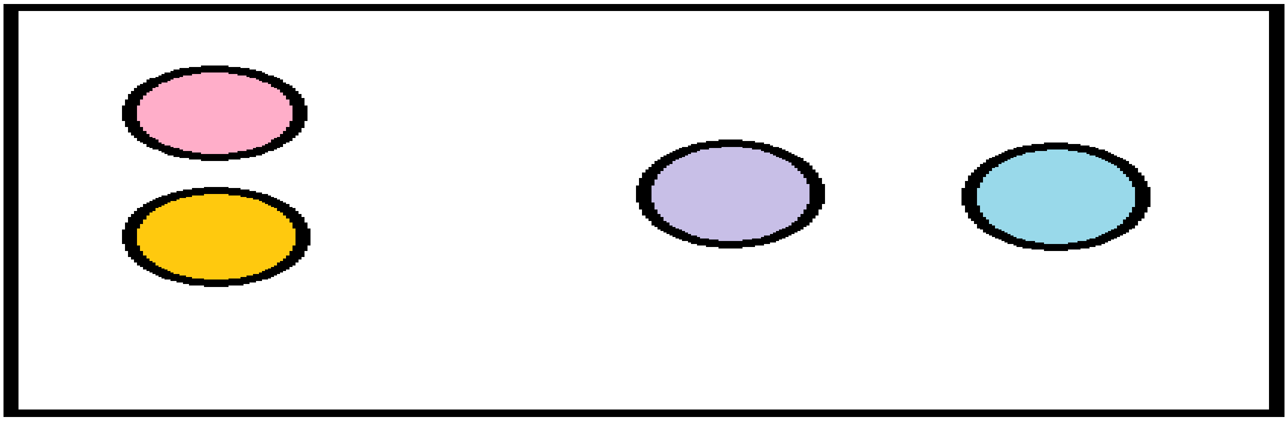


Fig.1.7 Well Shaped Cluster [19]

5. **Centre Based Cluster**

A cluster is a set of objects. An object in cluster is more close to the central of a cluster which is similar not to the centre of any other cluster. A centroid which is average of all points in cluster or a medoids which is most representative point in cluster and often the centre of a cluster.



Fig, 1.8 Centre based cluster [19]

1.5 K-Means cluster rule

The k-means cluster rule is that the basic rule that relies on partitioning technique that is employed for several cluster tasks particularly with low dimension datasets. It uses k as a parameter, divide n objects into k clusters in order that the objects within the same cluster area unit the same as alternative one another} however dissimilar to different objects in other clusters. The rule tries to seek out the cluster centres, (C1 …… Ck), such the total of the square distances of every datum, xi , 1 ≤ i ≤ n, to its nearest cluster centre Cj, 1 ≤ j ≤ k, is decreased . First, the rule arbitrarily selects the k objects, every of that at first represents a cluster mean or centre. Then, every object xi within the information set is assigned to the closest cluster centre i.e. to the foremost similar centre. The rule then computes the new mean {for every for every} cluster and reassigns each object to the closest new centre [24]. This method iterates till no changes occur to the assignment of objects. The convergence leads to minimizing the sum-of-squares error that's outlined because the summation of the square distances from every object to its cluster centre [3, 8]. The subsequent procedure summarizes the k-means algorithms [23]:

Algorithm: k-means:- The k-means rule for partitioning, wherever every centre of cluster delineated by the mean of the objects within the cluster.

Input:

k: the quantity of the cluster,

D: Knowledge containing the set of n objects.

Output:

A set of k clusters.

Method:

(1) Arbitrarily opt for k objects from D because the initial cluster centres;

(2) Repeat

(3) (re)assign every object to the cluster to that the article is that the most similar, supported the mean of the objects within the cluster;

(4) Update the cluster means that, i.e., calculate the mean of the objects for every cluster;

(5) till no change;

Despite getting used during a big selection of applications, the k-means rule has following drawbacks [25]:

1. As several cluster ways, the k-means rule assumes that the quantity of clusters k within the info is understood beforehand that, obviously, isn't essentially true in real-world applications.

2. As associate repetitious technique, the k-means rule is very sensitive to initial centres choice.

3. The k-means rule could converge to native minima.

1.5.1 Genetic Algorithm:-

Genetic rule originated from the studies of cellular automata. It’s conducted by John European nation and his colleagues. A Genetic rule is essentially a looking techniques, it's employed in the pc science. It helps to seek out approximate solutions for any improvement issues. The genetic algorithms area unit referred to as the biological process algorithms. During this several techniques area unit concerned by biological process biology like inheritance, mutation, natural process, and recombination. Within the illustration of the genetic algorithms the fitness performs is outlined [11]. The genetic rule payoff to initialize the solutions arbitrarily. It won’t to improve it through repetitive application. during this case it involves several applications such as: mutation, crossover, and choice operators. Several Researchers have adopted genetic algorithms as an answer to improvement in numerous fields. The genetic algorithms acts as an answer to improvement downside started gaining quality towards the top of the last century as wont to solve improvement issues in construction. Its intrinsic correspondence facilitates the uses of distributed process machines, like Distribution Network designing. Issues that seem to be significantly acceptable for answer by GA embrace programming and State Assignment downside. GA approach to unravel Map colour downside has been examined conjointly. Researchers have shown interest in GA approach to unravel programming varieties of issues, like job look programming downside. It is often quite effective to mix GA with different improvement ways. Hybrid GA approach is additionally being adopted to derive higher quality solutions in comparatively shorter time for exhausting combinatorial universe improvement issues like bagman downside (TSP). Of late, analysisers are attempting to explore the facility of GA in numerous field of research like Molecular analysis, genetic analysis to spot unknown genes of comparable perform from expression information. The GA could be a random world search technique that mimics the figure of natural biological evolution. Genetic algorithms treat a population of potential solutions. During this numerous principle of survival area unit applies. This principle area unit helps for fittest to provide higher approximations to an answer. Within the genetic algorithms, the degree of downside domains area unit outlined. Some operator's area unit borrowed from natural genetic science. This method wants the evolution of populations of people. This area unit higher suited to their atmosphere than the people. The foremost ordinarily used illustration in genetic algorithms is that the binary alphabet though different representations are often used, e.g. ternary, integer, real-valued etc. as an example, supposes there are a unit 2 variables X1 and X2. In these variables there's downside. This downside is often mapped within the following structure within the following way: Here, X1 containing ten bits whereas the X2 containing the15 bits. It always reflects the amount of accuracy of the individual call variables. During this case, if the developer examines it on the premise of isolation, he or she might not found any helpful data. The knowledge is often fetching solely by decrypt the body. It's solely with the cryptography of the body into its constitution values that any that means are often applied to the illustration. It's following advantages:

1. It is an optimization technique.
2. Its applicability on the wider range problem.
3. It is also work with population of solution.
4. The genetic operator in helps to produce new population.

1.5.2 Cluster Analysis

Cluster analysis may be define as a creating group of object in such a way that the object in a group will be related to one other and un related to the object in another group. We can’t say that cluster analysis is not restricting to one specific algorithm, but its main motive to solve the task. It can be also achieved by using algorithm but in different manner. Clustering analysis used in various felids hat is biology, trade and business. In each of the field clustering analysis has is on importance. In broader view we can say that clustering analysis is not limited to computer team as it is contributing to other felids also.

Chapter-2

REVIEW OF LITERATURE

Shaheda Akthar1, Sk.Md.Rafi , "Improving the Software Architecture through Fuzzy Clustering Technique"

In this paper they explained about the reverse engineering concept is quite famous these days and related to recovery of software architecture. There are number of technique which as used in this paper to recover software architecture, one of them is clustering technique, which source the same component from software. Generally the component feature is vague. A group of same data element is known as clustering. This technique is as older and its used also in science and engineering. In simple words, identifying he number of data element, calculating similar coefficient and following the clustering method is called as clustering technique. The main function of the clustering technique for speedy and efficient recovery of software architecture by using fuzzy clustering technique. In this paper the major impact of this study shown that architecture recovery can be done batter by fuzzy clustering instead of ordinary clustering. [9]

Chih-Cheng Hung!, Wenping Liu and Bor-Chen Kuo, "**A new Adaptive fuzzy**

Clustering algorithm for remotely sensed images"

In this paper they explained the adaptive fuzzy algorithm which is come along with the capability and adaptation. This adaptive caliber can be fulfill by using the tool of partition and consolidate it. The number of classes is the data set which requires the prior knowledge in fuzzy clustering algorithm. This new technique of algorithm can able to learn the number of classes continuously. Fuzzy mathematic provides the great accuracy results in clustering. The various techniques like k-mean, ISODATA, fuzzy C-mean and possibilistic C-mean algorithm is very effective where we require image segmentation. K-mean clustering identify the number of cluster continuously. The C-mean clustering and fuzzy C-mean clustering and new fuzzy clustering algorithm have an advantage when it combined with ISODATA. [10]

WANG Jing1, TANG Jilong, "Alternative Fuzzy cluster segmentation of remote sensing images based on adaptive genetic algorithm"

In this paper they generate the idea about a technique which is based on image understanding and its analysis is called as remote sensing image segmentation. This paper is introduce the image analysis which required various technique i.e. Adaptive Genetic Algorithm (AGA) and alternative fuzzy C-Mean. The AGA identified the segmentation. The remote sensing images are always difficult because of they are equal grey pixel may be divide into different region of clustering. It is the batter technique then the old technique which takes huge number of second. Whereas it take only needs few second. The segmentation process is the widely used technique in remote sensing images, which collect information, process of information and analysis. [11]

Markus Bauer Forschungszentrum Informatik Karlsruhe, "**Architecture-Aware Adaptive Clustering of OO Systems"**

In this paper they explained about Re-engineering software system is the recovery of software architecture and in software architecture recovery involves clustering. In this paper they guide us to introduce an approach that collectively clustering with matching technique to discover a decomposition which is well understood. Pattern matching is a technique under which architectural clues can be identified. All these clues are helpful to access an interclass similarity measure in clustering algorithm to produce the decomposition which is also known as final system decomposition. Adding a new updating in current existing software is always a challenging task but it also helpful to reduce the complexity in work. It is also necessary to keep update every error, patch or hack, for batter performance of any software system or a software architecture. Architectural clue collect the source model is designed with proper information. [12]

Zhang Chen et.al, “A Robust Fuzzy Kernel Clustering Algorithm”

In this paper they introduced that with a kernel function the classical Fuzzy kernel clustering method does iterative clustering in the actual data space or in the main space by mapping the samples into good dimension space. This method is used weak robustness against noises and outliers. The robust kernel clustering algorithm is a technique which represent to improve the robustness by using parameter. Possiblistic C-Means (PCM) is used to enhance the robustness using the typical parameters and easily produce consistent clustering. The PFCM is the combination of FCM and PCM which solve the consistency clustering and noise sensitive problems. The optimized kernel parameter is used to do possiblistic fuzzy kernel clustering. The FKCM and FKCO algorithm are sensitive and poor clustering effect. The parameter of kernel function optimization method under the condition which is not supervised which is presented and compared with the other similar kernel clustering algorithm, this algorithm can not only deal with the linearly inseparable dataset, and can achieve great clustering efficiency through noise jamming. [13]

D. Doval, S. Mancoridis, B. S. Mitchell, “Automatic Clustering of Software Systems using a Genetic Algorithm”

In this paper give us a brief idea about large software system to have a complicated structure. Software designer usually illustrate the structure of software system as one or more graph which is called directed graph. Directed graph explain the classes of a system and their relationships using various nodes and directed edges, is called as module dependency graph. Module Dependency graph (MDG) can be big and complicated graph. MDG partition is a technique which is well defined in this paper. Good partition characteristic usually independent subsystem that accumulate which are interdependency. To discover a batter partition as an optimization problem and user a Genetic Algorithm (GA) to find the distinct large solution space of all possible MDG partition, improve the performance of GA input encoding, various genetic operator. [14]

Narayan Desai, Rick Bradshaw, Ewing Lusk, "**Component-Based Cluster Systems Software Architecture a Case Study"**

In this paper, they identify that the component architecture in an idea of an approach which is not been traditionally applied in any part of cluster system software. According to this paper cluster system, the group of program used to construct and manage individual node, together with the software involved in submission, scheduling, monitoring and termination of jobs. Component architecture have wide range of helping property for developer and user, it also describe how the component approach maps in system software problem, with the experiences and with the approach in implementing an all new suit of system software for small or medium sized cluster with normally complex system software needs. The difficulty of the system software has reduced while sharpness and reliability have improved. Component architecture in system software development is a encouraging methodology for both developer and user. By using this approach increase a great number of component implementation will become possible. [15]

Chung-Horng Lung, "Software Architecture Recovery and Restructuring through Clustering Techniques"

In this paper they proposed the software architecture is analytic for the maintenance and changes in any software system. The ongoing approaches are restricted to software architecture recovery in the reverse engineering. This paper also gives us idea to present a qualitative approach which is based on clustering models for software architecture reconstructing and re-engineering and software architecture recovery. Clustering technique is one of the important reverse engineering tool. We have two case studies in this paper. First is experimental study of a decoupling creation for the real time telecommunication system. The second one present a study on a possible to accomplish the user to improve architecture by using a design based on a clustering outcome. An approach depends upon the clustering technique not to discover software architecture, but also enhance for batter result. Use cases are use simultaneously along with the clustering technique for less complication. The visualization tool SPV (Software partition and visualization) is helpful to create a user friendly environment by using various clustering techniques. [16]

Ioana S¸ ora, Gabriel Glodean, Mihai Gligor,” Software Architecture Reconstruction an Approach Based on Combining Graph Clustering and Partitioning”

In this paper present an idea to enhance the efficiency of an automatic software architecture reconstruction. Most of the research gives us the idea about clustering direction in architecture reconstruction. The software architecture of a program is the basic design or designs of the system which contain different software elements, the major properties of those elements and correlate among each other. By studying this paper we also come across with a view that the result of coupling/cohesion driven clustering by linking it with the splitting preprocessing that build a layer of the different classes of the system. Clustering algorithm is widely use in the data mining to describe class of objects whose members are related among each other from some way. In reverse software engineering clustering help to generate architectural application, sub system modules grouping among each other (classes function etc). The portioning was explained in the development of the product process and it refer to rearrange the design function in order to increase the availability of information which is necessary at each level of the process. [17]

Kamran Sartipi ,”Software Architecture Recovery based on Pattern Matching”

In this paper presents the technique and its relevancy for discovering the high level design of legacy software system depends upon on user well explained architectural system and another technique which is called graph matching technique. After studying this paper we are well versed with its proposed model, and high level point of view related to software system by representing as its query, using a detail language. Pattern graph is helping out in mapping of a query .On the other hand graph-nodes and a group of graph edges presents the interaction with among the component. Interaction constrain can be resented by the detail language as a query. Pattern/graph in used against the entity relation graph which represents the required information from the source code of the software system. Pattern based model initially create the high level mental model of the system architecture, from this modeling technique help software architecture identify the query language and a matching engine searches in software system. [18]

Maninderjit Kaur and Sushil Kumar garg, "Survey on Clustering Techniques in Data Mining for Software Engineering", 2014

In paper presents that Quality and reliability of the computer software is very important. Software development uses a huge amount of software engineering data. Software Engineering data is the collection of execution traces, code bases, graphs, bug reports etc. Software Engineering data is very useful in understanding the development and working of any product or software. Software is of high quality and highly reliable if it is error-free. Software is error-free if there is no bug present in it or it is free from bugs. Bugs are very hard to find. Software Engineering tasks are Programming, Testing, Bug Detection, Debugging and Maintenance. Data Mining Techniques are applied on software engineering tasks. Data mining techniques are used to mine software engineering data and extract the meaningful and useful information. Techniques used for mining software engineering data are matching, clustering, classification etc. Every technique has to solve different problems and have their own advantages and disadvantages. There is no such clustering technique and algorithm exists that is used to solve all the problems and is a best fit for all applications. As the application changes requirements will also change. With this change the selection of clustering technique affected. No technique or algorithm is the readymade solution to all applications and problems. Predefined number of clusters and stopping criteria affect the accuracy and performance of clustering. Handling of noisy data, data set size, shape of the clusters all affects the clustering results. Based on the application, they have to choose the suitable clustering technique and algorithm in future work. [19]

Manpreet Kaur and Usvir Kaur, "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", 2013

In this paper they introduced question redirection provides a mechanism for B1 Server to work out the set of logical table sources (LTS) applicable to a logical request whenever missive of invitation will be glad by over one LTS. The Oracle B 1 repository shipped in Oracle Fusion applications contains metadata information content for time period coverage analysis (using Transactional Business Intelligence) and historical coverage (using B1 Applications).The planned work represents question redirection methodology that improved K-means agglomeration algorithmic rule performance and accuracy in distributed atmosphere. During this paper they need done analysis on k-mean and stratified algorithmic rule by applying validation measures like entropy, f-measure, constant of variance and time. The experimental results show that k-mean algorithmic rule performs higher as compared to stratified algorithmic rule and takes less time for execution. [20]

Tapas Kanungo and Nathan S. Netanyahu, "An Efficient k-Means Clustering Algorithm:Analysis and Implementation", 2002

In this paper they introduce k-means clustering, they're given a collection of n knowledge points in d-dimensional space Rd Associate in an integer k and also the downside is to see a collection of k points in Rd, known as centres, therefore on minimize the mean square distance from every information to its nearest centre. a well-liked heuristic for k-means clump is Lloyd's algorithmic program. during this paper, we tend to present an easy and economical implementation of Lloyd'sk-means clustering algorithmic program, that we tend to decision the filtering algorithmic program. This algorithmic program is straightforward to implement, requiring a kd-tree because the solely major organisation. They establish the sensible potency of the filtering algorithmic program in 2 ways that. First, they present a data-sensitive analysis of the algorithm's time period, that shows that the algorithmic program runs quicker because the separation between clusters will increase. Second, they gift variety of empirical studies each on synthetically generated knowledge and on real knowledge sets from applications in color quantisation, knowledge compression, and image segmentation. [21]

Amar Singh and Navjot Kaur, "To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm", 2013

In this paper work represents ranking primarily based methodology that improved K-means clustering algorithmic program performance and accuracy. during this they have additionally done analysis of K-means clump algorithmic program, one is that the existing K-means clustering approach that is incorporated with some threshold price and other is ranking methodology that is weighted page ranking applied on K-means algorithmic program, in weighted page rank algorithmic program chiefly in links and out links ar used and additionally compared the performance in terms of execution time of clustering. planned ranking primarily based K-means algorithmic program produces higher results than that of the present k-means algorithmic program. [22]

K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and potency of the k-means agglomeration Algorithm", 2009

During this paper they introduced emergence of recent techniques for scientific information assortment has resulted in giant scale accumulation of information referring to numerous fields. typical information querying ways square measure inadequate to extract helpful info from vast information banks. Cluster analysis is one in all the key information analysis ways and therefore the k-means agglomeration rule is wide used for several sensible applications. however the initial k-means rule is computationally pricey and therefore the quality of the ensuing clusters heavily depends on the choice of initial center of mass. many ways are planned within the literature for rising the performance of the k-means agglomeration rule. This paper proposes a way for creating the rule more practical and efficient; therefore on regain agglomeration with reduced quality. [23]

Shuigeng Zhou et.al, "A neighborhood algorithm", 2010

During this paper they explained that the foremost necessary ways for data discovery in databases (KDD), agglomeration is incredibly helpful in several application areas, like alpha information analysis, business intelligence, and image process etc. Currently, there square measure principally four varieties of agglomeration algorithms: partitioning, class-conscious, density-based and grid-based algorithms. During this paper, we have a tendency to gift a brand new clustering rule, NBC, i.e. Neighborhood based mostly agglomeration that discovers clusters supported the neighborhood characteristics of information. The NBC rule has the subsequent advantages: (1) NBC is effective in discovering clusters of arbitrary form and completely different densities; (2) NBC wants fewer input parameters than the present agglomeration algorithms; (3) NBC will cluster each giant and high-dimensional databases with efficiency. [24]

Chapter-3

SCOPE OF STUDY

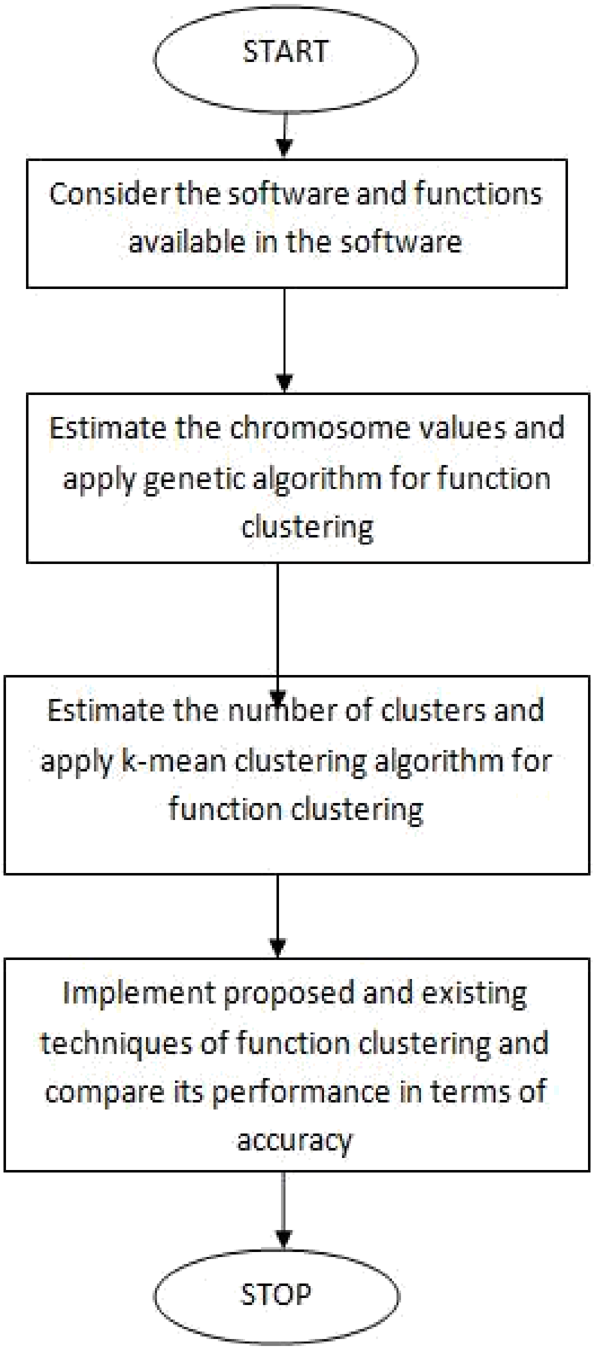
The complexness and therefore the nature of the computer program had modified significantly within the last thirty years. The previous applications will run on single processor and generate permanent or fixed output .But with the development within the technology application area unit having the troublesome the computer program and these applications run on the varied systems along like applications that that support the design of shopper server. gift day applications will run on varied in operation systems as a result of the complexness and therefore the nature of the applications we'd like to calculate the performance and different application issue .To calculate the applying performance we wish to characterize some set of rules. Therefore, we have a tendency to approve the construct, ways and practices of the package engineering. With the employment of the package engineering ideas and therefore the ways of software engineering we are able to the performance of applications and therefore the different elements. We’ve got to investigate a number of main failures that may ends up in failure of software before convincing the user applications. The software failing complication can be raised in the complex software’s, when we are not able to properly analyze the properties of the software. In the past times the algorithm of genetic had been proposed to cluster the functions of similar properties. In the genetic algorithms, all the clustering values are depends on the chromosomes. It is very difficult to estimate the correct value of chromosomes, which decreases the efficiency of the software architecture analysis. For increasing the software architecture analysis, the K-MEAN clustering will be used which is more efficient then the genetic clustering. This will improve the software architecture analysis and improve the accuracy and reduce algorithm escape time.

Chapter-4

OBJECTIVE OF THE STUDY

4.1. PROBLEM FORMULATION

The software architecture contains number of functions and modules. Among all the functions in the software some functions are necessary and some are not according to their importance and functionality. To properly classify the categories of the modules given approach had been suggested in the earlier times among all the suggested approach clustering is the most efficient technique for clustering the similar type of functions. In the base paper, genetic algorithms had been applied for clustering the similar type of data. The algorithms of genetic are depend on the chromosome values, which is the inefficient technique of clustering. When the genetic algorithm is applied to cluster similar type of functions, the accuracy of function cluster will be reduced .As some functions are clustered in into important functions and other functions are clustered into non important function. To increase accuracy of the function clustering new technique will be proposed to efficiently cluster the functions according to their importance. To cluster same type of functions as they are valuable or not, the clustering K-mean algorithm will be applied. The K-mean algorithm will efficiently cluster the functions according to their importance because in k-mean clustering we know number of functions in the software and according to that we can define number of clusters for k-mean clustering.



START
Consider the software and functions available in the software
Estimate the chromosome values and apply genetic algorithm for function clustering
Estimate the number of clusters and apply k-mean clustering algorithm for function clustering
Implement proposed and existing techniques of function clustering and compare its performance in terms of accuracy
STOP

Fig. 4.1 Problem Formulation

4.2. OBJECTIVES

1. To study and analyze various software architecture analyze techniques in system software.

2. To identify the problem of accuracy in genetic algorithm for clustering of similar type of functions.

3. To propose new technique of clustering to cluster the similar type of functions to analyze the architecture of similar type of software's.

4. To implement recommended and current algorithms and analyze the accuracy and escape time performance.

Chapter-5

RESEARCH METHODOLOGY

We will use the K-mean clustering to cluster similar type of function in software architecture. The algorithm of K-means clustering algorithm is the main algorithm which is based on method of partitioning which is used for many clustering functions specifically which dimension has low datasets. The k is uses as a parameter, divide n objects into k clusters so that the same cluster objects are identical to each other but not identical to other objects in other clusters. The algorithm tries to search the centre of cluster, (C1 ...... Ck), such that the sum of the squared distances of each data point, xi, $1 \leq i \leq n$, to its nearest cluster centre Cj, $1 \leq j \leq k$, is minimized. First step, to selects the algorithm of the k objects randomly, each of which at first presents a mean of cluster or centre of the cluster. Then, the all object of xi in the set of data is selected for the closest centre of the cluster i.e. to the most similar centre. The algorithm then calculates the new mean for every cluster and reassigns every object to the closest new centre. This process iterates before no any kind of alternations occur to the objects of the assignment. The concur conclusion in reduce the sum-of-squares error which are defined as the detail of the squared distances from each and every object to its centre of cluster. The following summarizes the k-means algorithms. This algorithm will cluster the same type of functions and we will analyze the same type of functions. The K-means clustering is one of the understandable freely learning algorithms that solve the clustering problems. For that the procedure follows a simple and easy way for analyzing the given data set with a certain number of fixed clusters a priori. We use Matlab tool for implementing the algorithm of k-Means clustering.

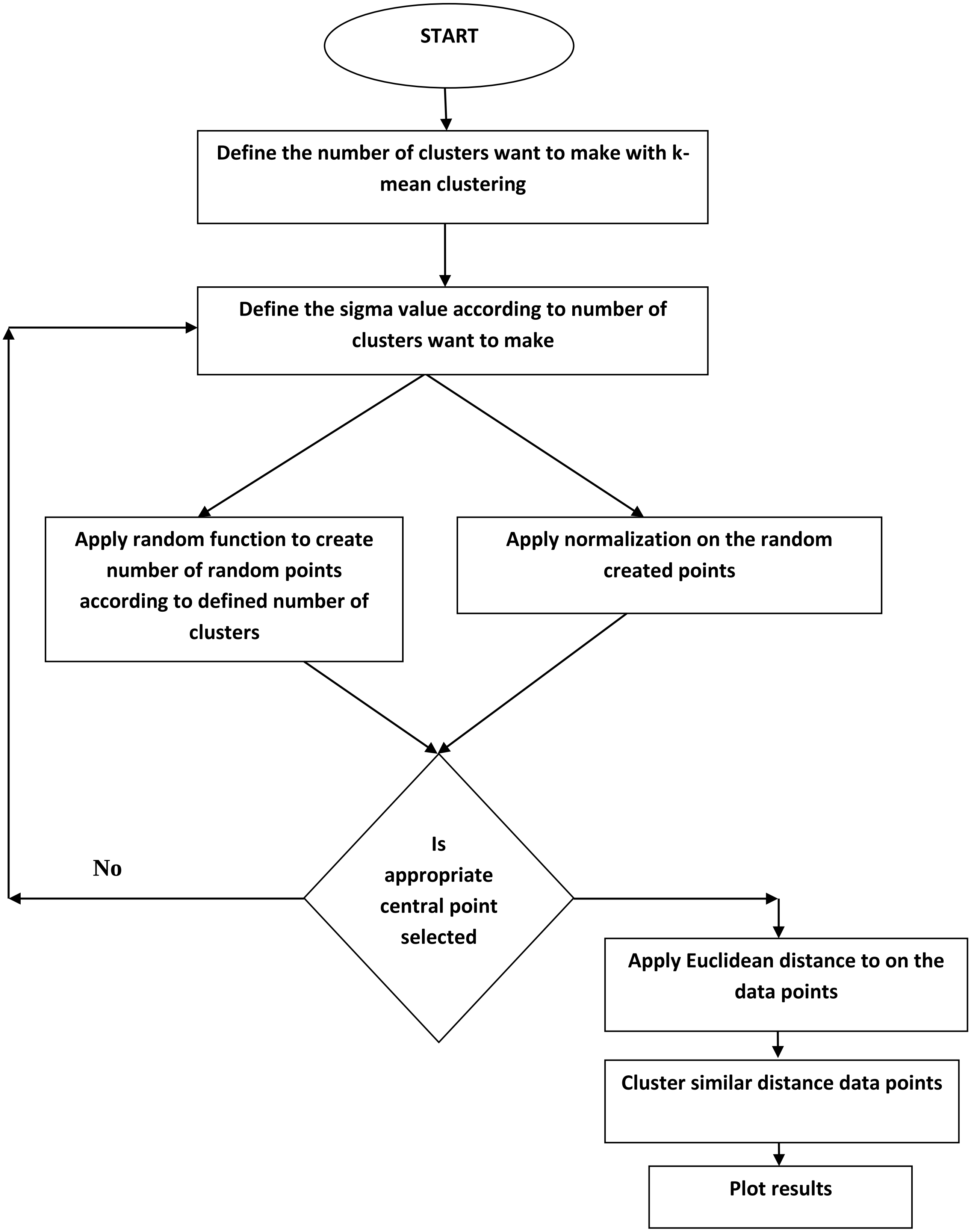

START
Define the number of clusters want to make with k-mean clustering
Define the sigma value according to number of clusters want to make
Apply random function to create number of random points according to defined number of clusters
Apply normalization on the random created points
Is appropriate central point selected
No
Apply Euclidean distance to on the data points
Cluster similar distance data points
Plot results

Fig. 5.1 Flowchart of proposed technique

As illustrated in the figure 1, the procedure of Meta clustering is shown. In this flowchart functions which are implemented works in following manner

1. **Load dataset and Sigma value Initialization:** In the first step of the flowchart, the dataset will be initialized on which the meta clustering has to be performed. After loading the dataset, the sigma function will be applied on the dataset. In the sigma function, the initial points are selected from the dataset on the bases of probability to select centroid for meta clustering
2. **Random function: - In** this step random function is applied to select point randomly from the dataset for the normalization. When normalization point is selected from the dataset.
3. **Unique and normalization: -** The dataset which is loaded firstly hierarchical clustering is performed, after calculating hierarchy of the dataset unique points are calculated from each hierarchy. The unique points are then normalized from hierarchical points
4. **Clustering: -** The point which are unique and are normalized, then the clustering is performed on 2 D plane. Each cluster is marked with different colours

Algorithm

```
    INPUT: Input data for clustering

      OUTPUT: Clustered Data

Start ()

      1. A=input data for clustering
      2. [a b]=size(A);
      3. Define number of clusters c;
      4. Define central point of the input data for clustering
      5. Apply normalization technique to redefine central point from input data
      6. Apply Euclidian distance to find nearest points from the input data
      7. Cluster the similar data points and plot the data

End ()
```

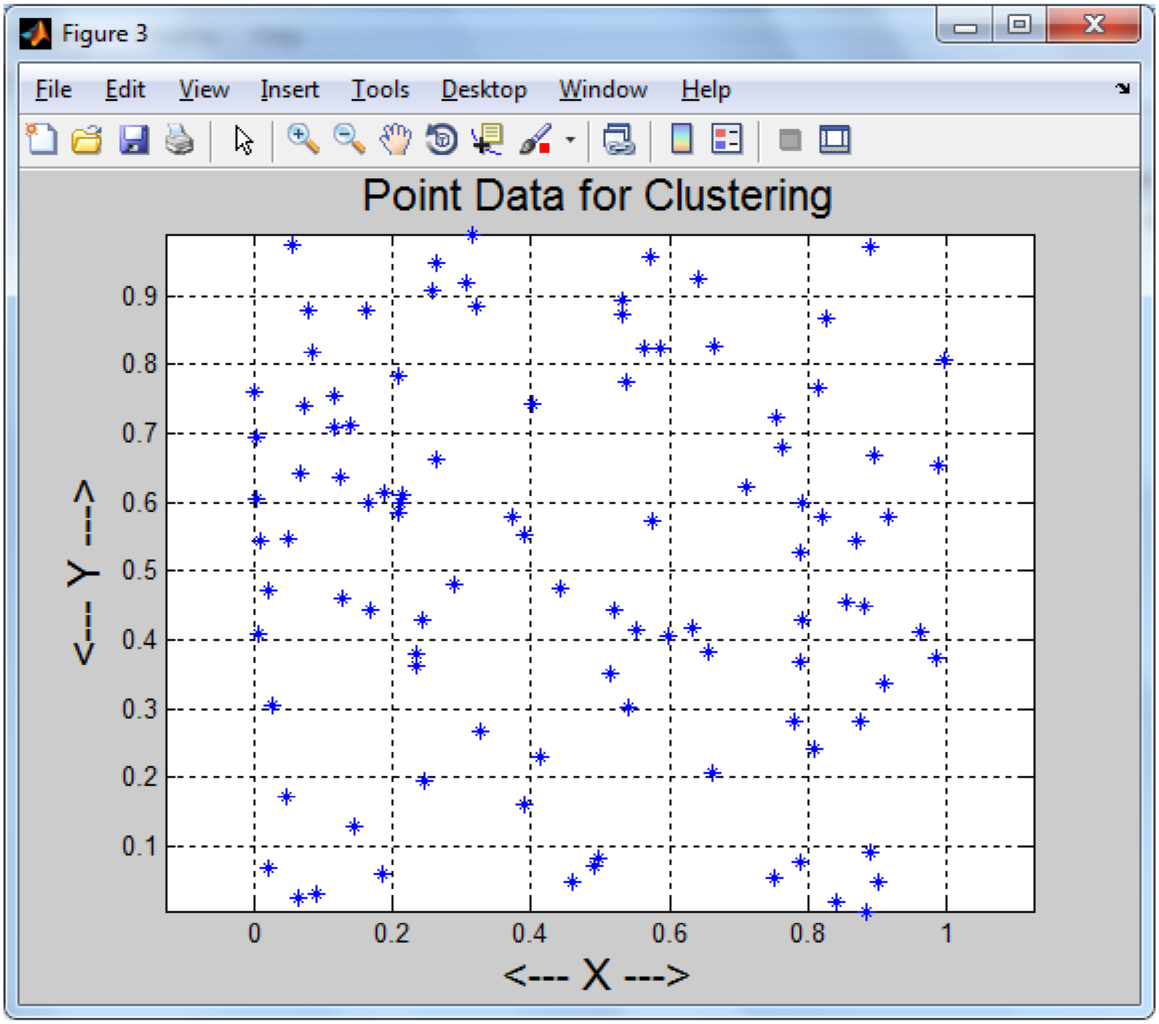


Fig. 6.1 K-mean clustering

As shown in figure 1, the dataset will considered and first of all the number of clusters will be defined after that central points are selected. The formula of Euclidian distance will be applied to cluster similar points

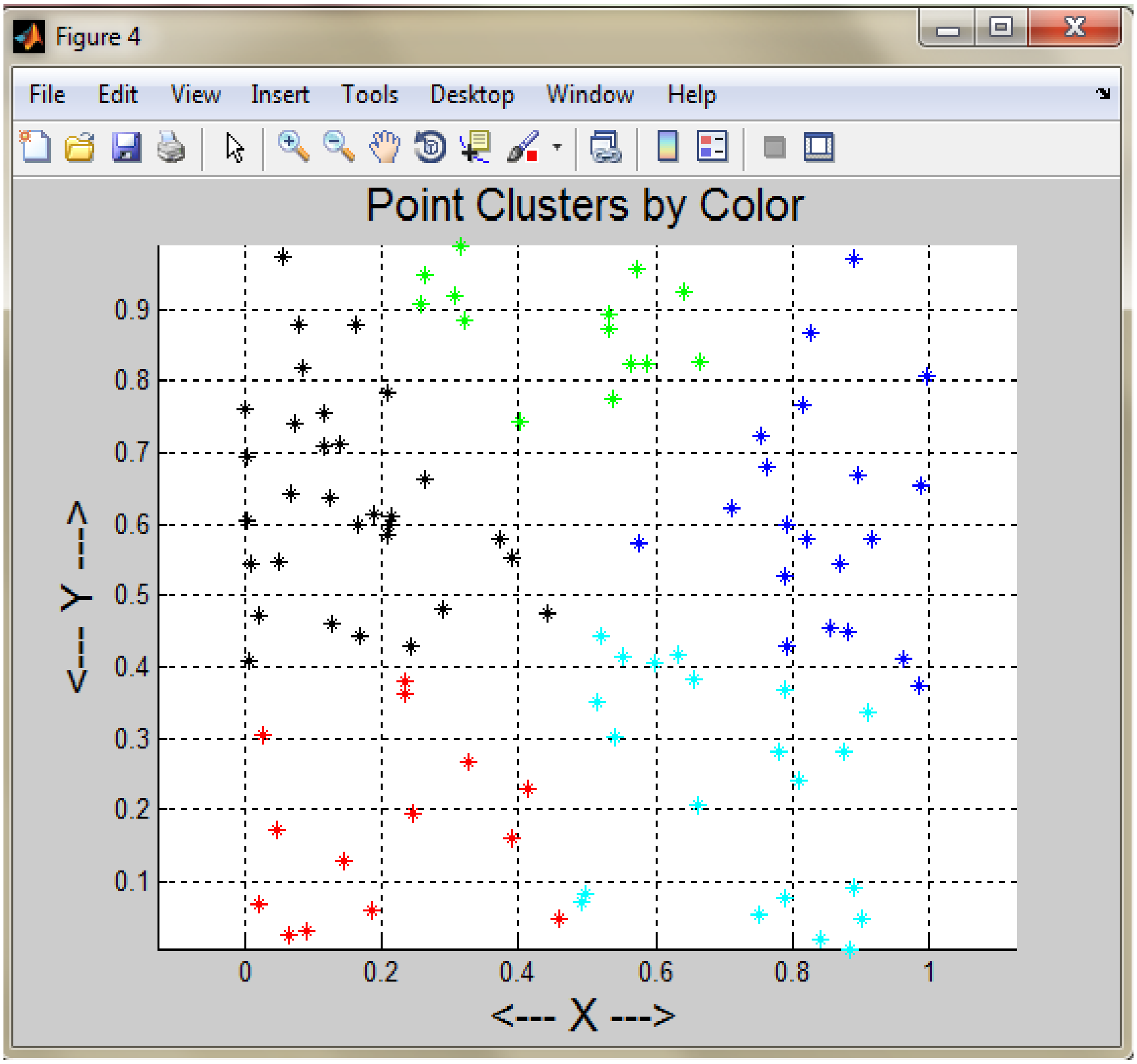


Fig.6.2 K-mean clustering

As shown in figure 2, the dataset will be clustered according to the different colors. The formula of Euclidian distance will be applied to cluster different point of cluster by colour

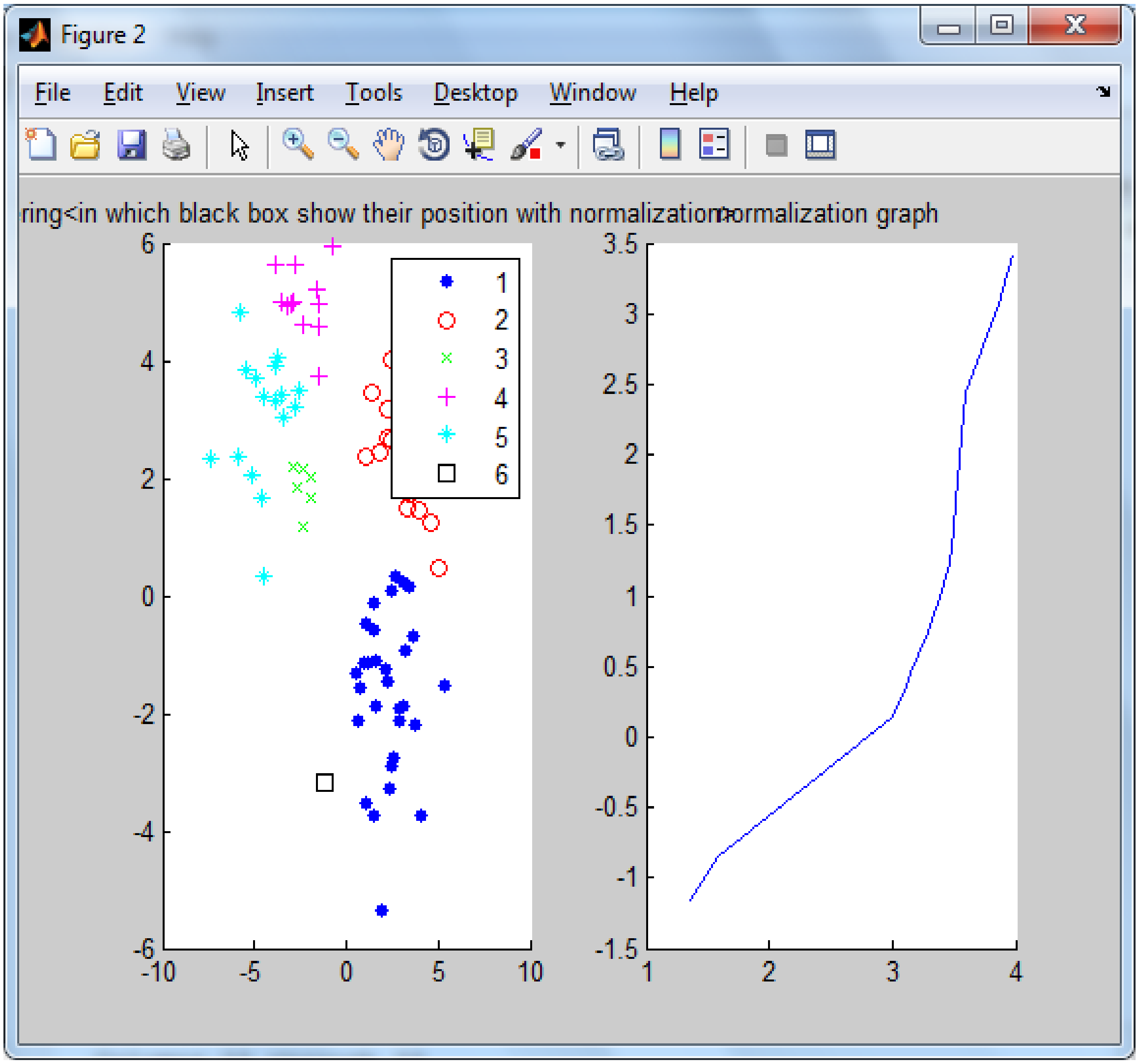
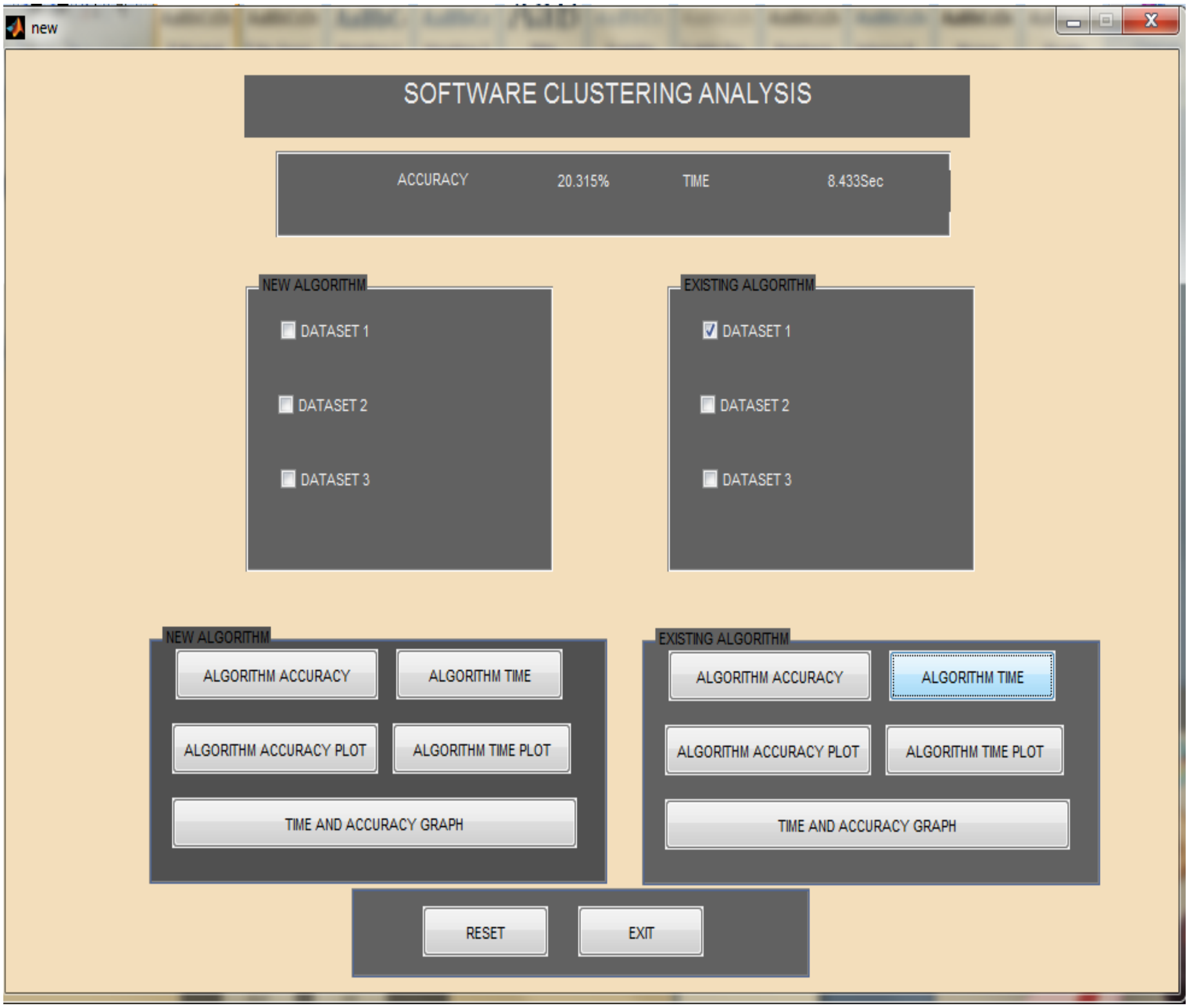


Fig. 6.3 Enhanced K-mean clustering algorithm

As shown in figure 3, the enhancement will be applied in k-mean clustering algorithm. In the enhanced K-mean clustering algorithm, dataset will be loaded. In the second step random central points is selected on the basis of probability function. Then normalization will be applied to select most relevant central point. After the selection of central point Euclidian distance will be applied to cluster similar points



ACCURACY 20.315% TIME 8.433Sec

NEW ALGORITHM

□ DATASET 1

□ DATASET 2

□ DATASET 3

EXISTING ALGORITHM

☑ DATASET 1

□ DATASET 2

□ DATASET 3

Fig. 6.4 Performance analysis

As shown in figure 4, interface is designed for the performance analysis of existing algorithm of k means and enhanced algorithm of k means. The performance analysis will be done in terms of accuracy and time. In this figure accuracy on Dataset1 is 20.315 % and time is 8.43 sec

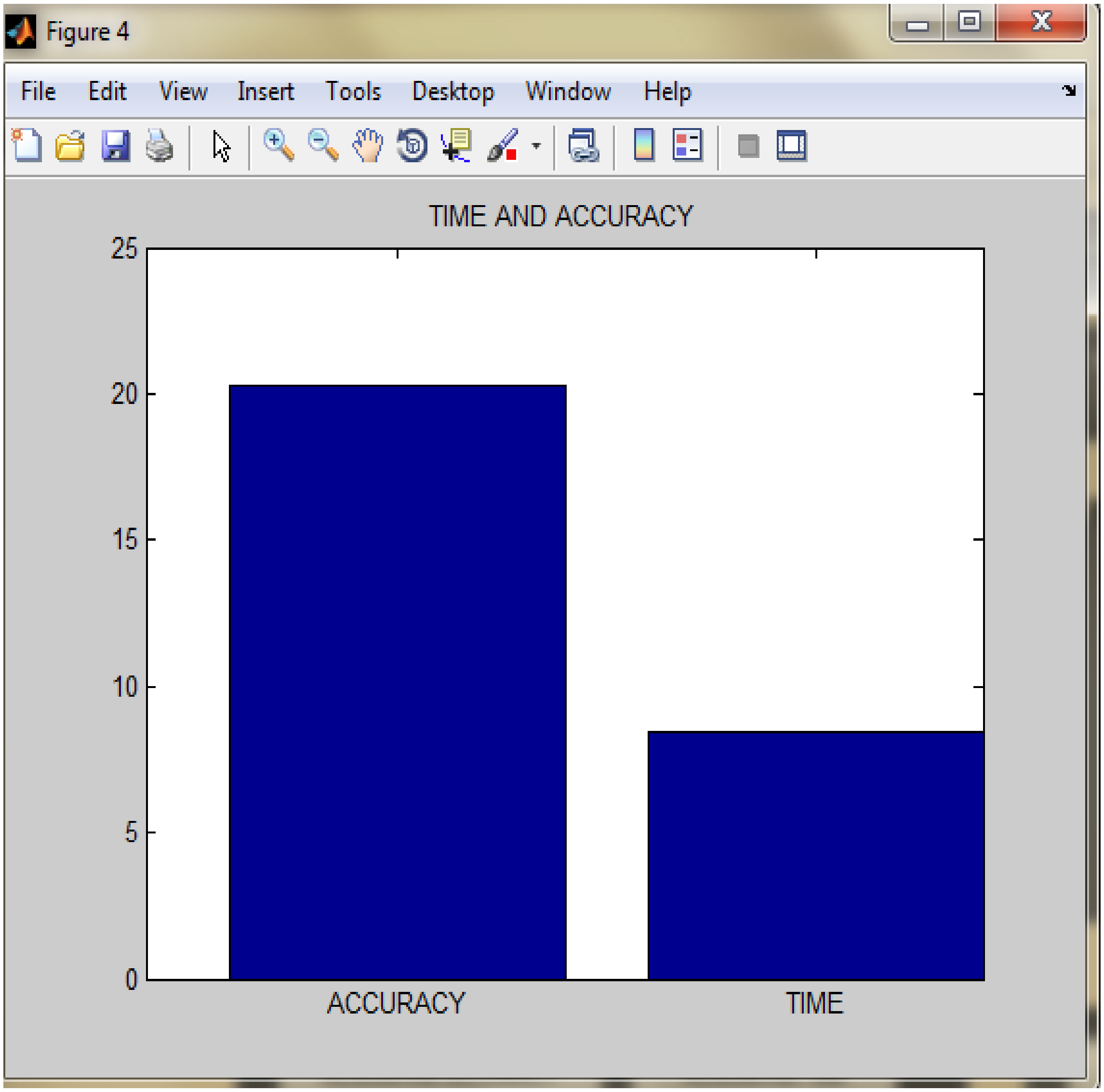
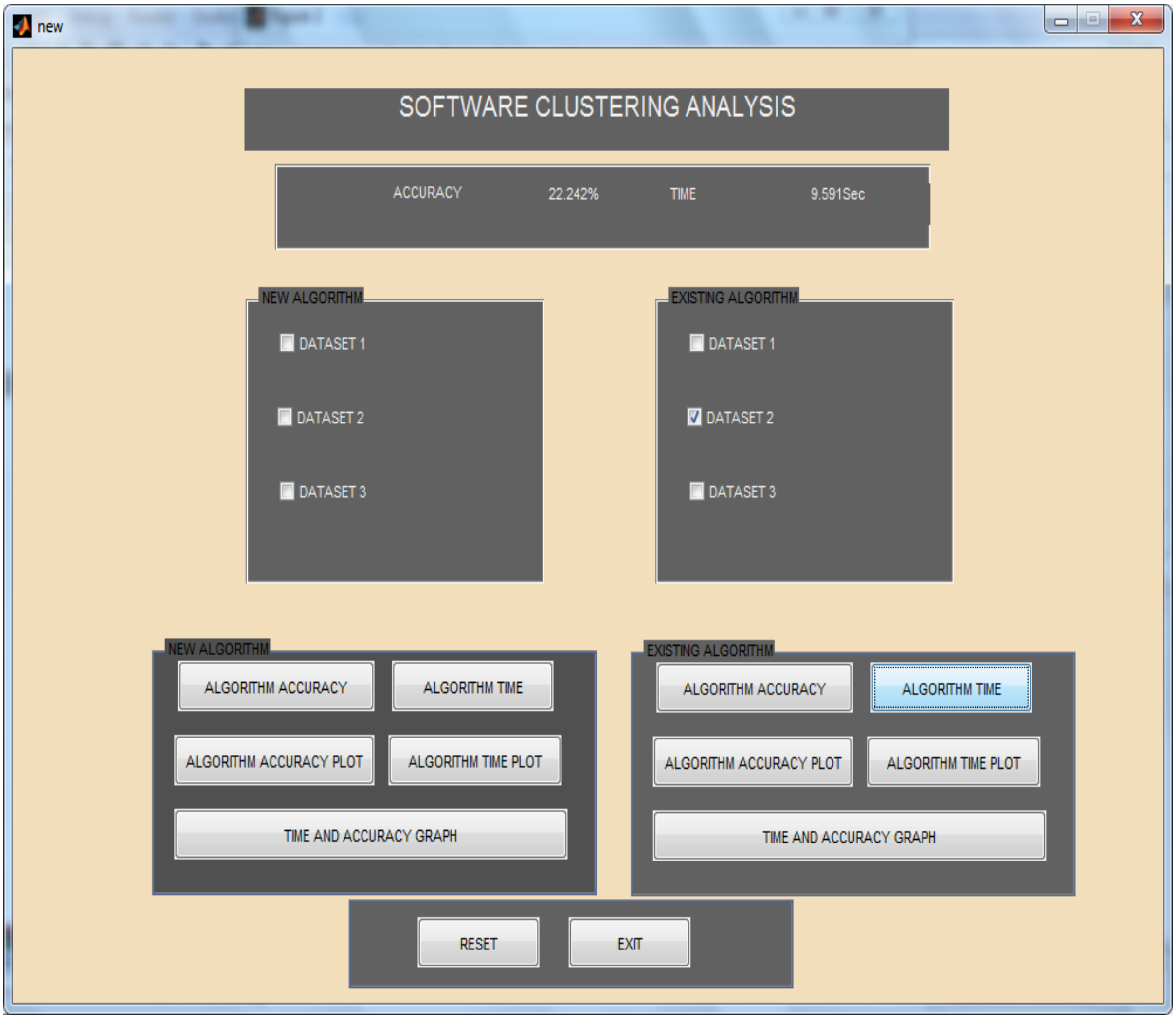


Fig. 6.5 Performance of existing algorithm

As illustrated in figure 5, the existing k-mean algorithm performance show in form of bar graph. The accuracy on dataset 1 is 20.315 is % and time is 8.433 sec



ACCURACY 22.242% TIME 9.591Sec

NEW ALGORITHM

□ DATASET 1

□ DATASET 2

□ DATASET 3

EXISTING ALGORITHM

□ DATASET 1

☑ DATASET 2

□ DATASET 3

Fig.6.6 Performance analysis

As shown in figure 6, interface is designed for the performance analysis of existing algorithm of k means and enhanced algorithm of k means. The performance analysis will be done in terms of accuracy and time of K-mean algorithm. In this figure accuracy on Dataset2 is 22.243 % and time is 9.59 sec

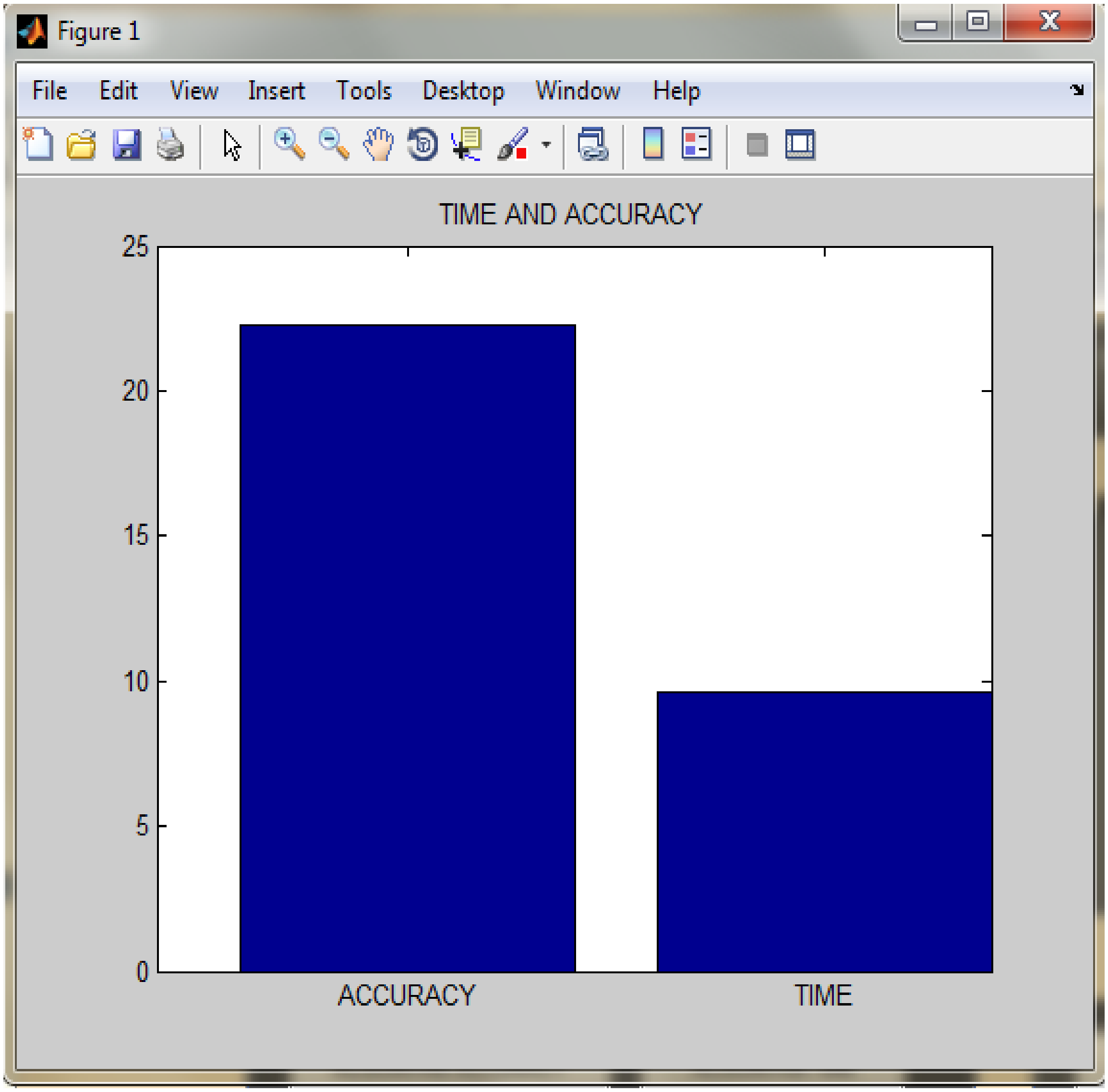
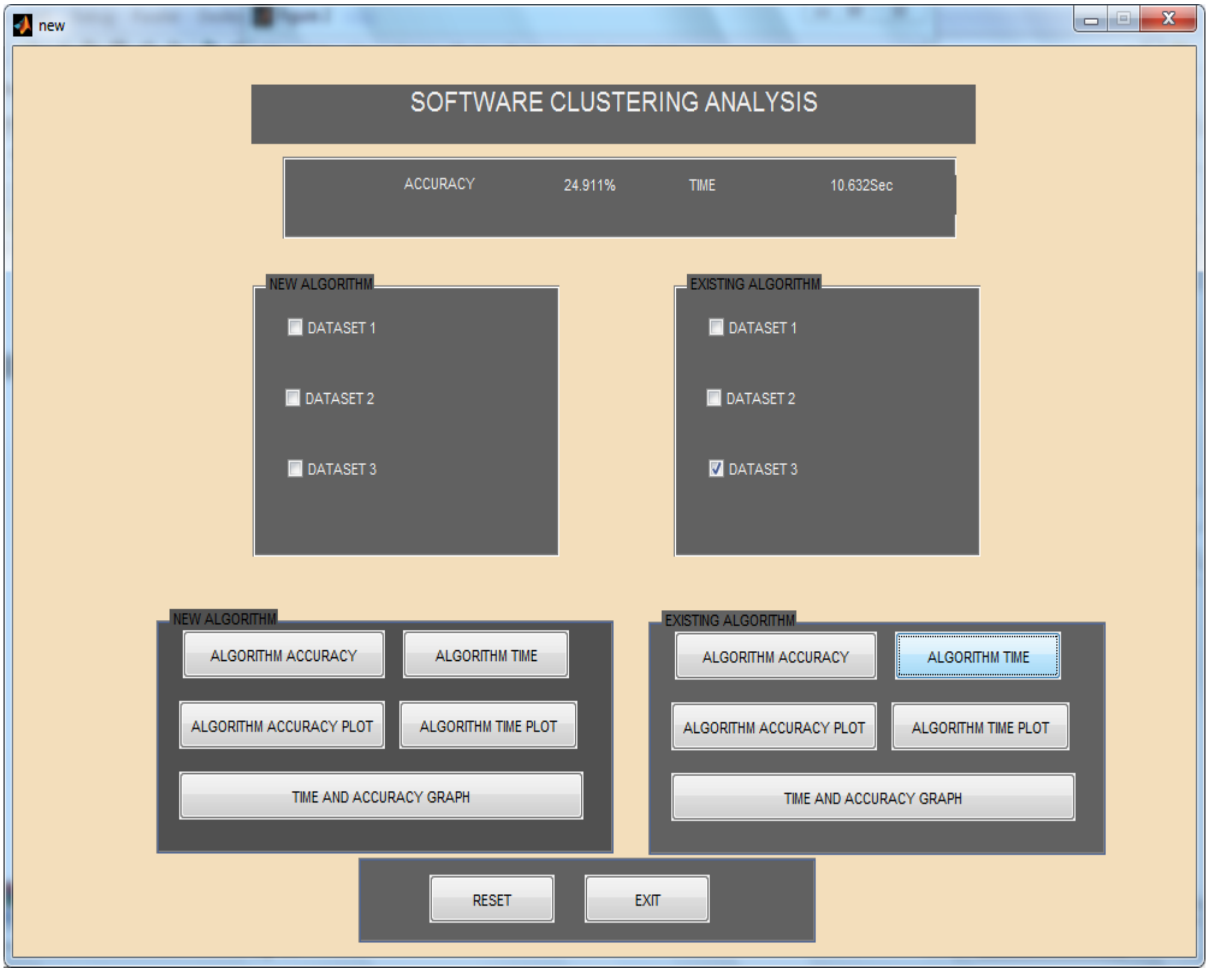


Fig.6.7 Performance of existing algorithm

As illustrated in figure 7, the existing k-mean algorithm is performance show in form of bar graph. The accuracy on dataset 2 is 22.242 is % and time is 9.591 sec



ACCURACY 24.911% TIME 10.632Sec

NEW ALGORITHM

□ DATASET 1

□ DATASET 2

□ DATASET 3

EXISTING ALGORITHM

□ DATASET 1

□ DATASET 2

☑ DATASET 3

Fig.6.8 Performance analysis

As shown in figure 8, interface is designed for the performance analysis of existing algorithm of k means and enhanced algorithm of k means. The performance analysis will be done in terms of accuracy and time of K-mean algorithm. In this figure accuracy on Dataset3 is 24.91 % and time is 10.59 sec

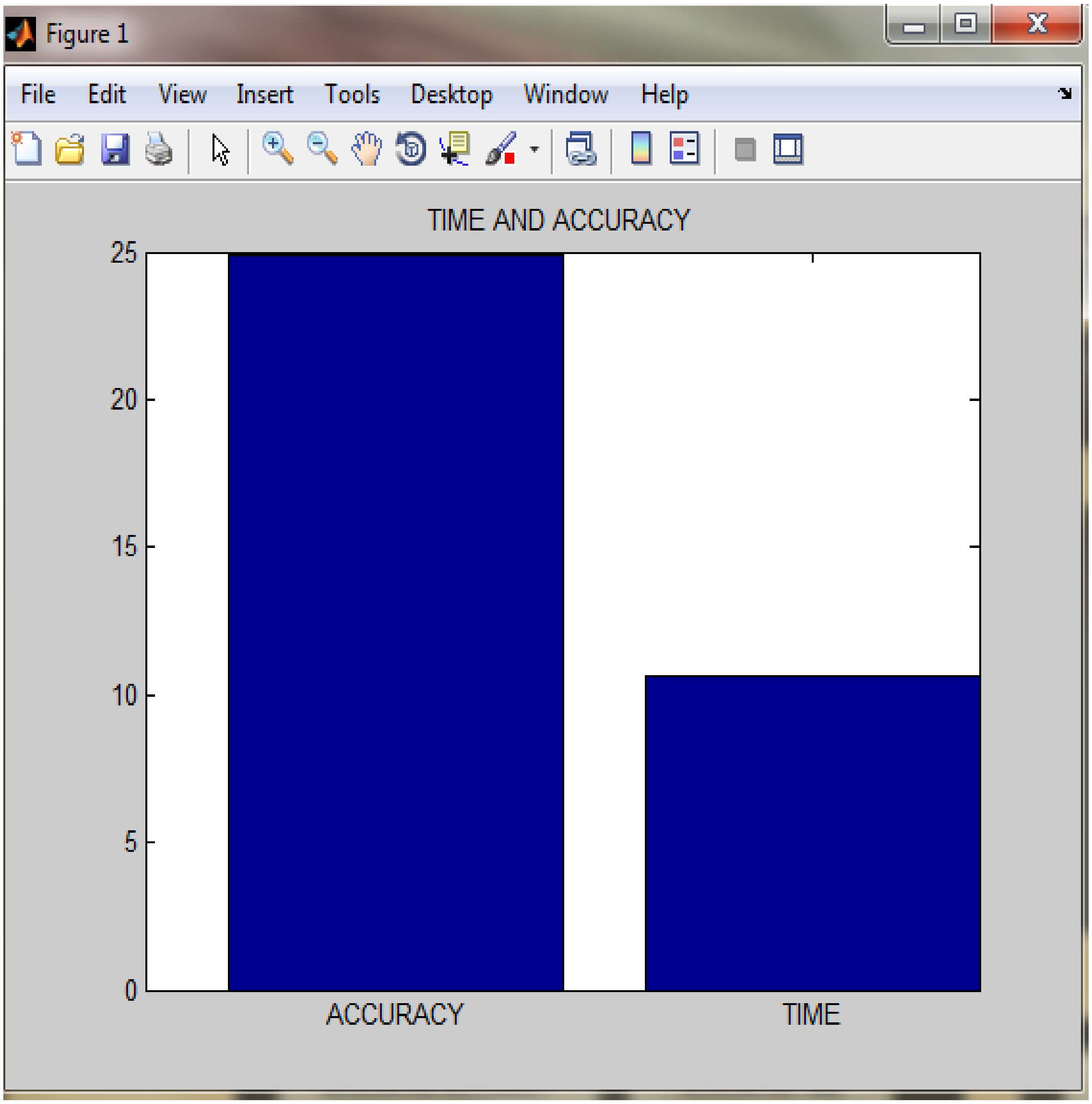


Fig. 6.9 Performance of existing algorithm

As illustrated in figure 9, the existing k-mean algorithm is performance show in form of bar graph. The accuracy on dataset 3 is 24.91 is % and time is 10.59 sec

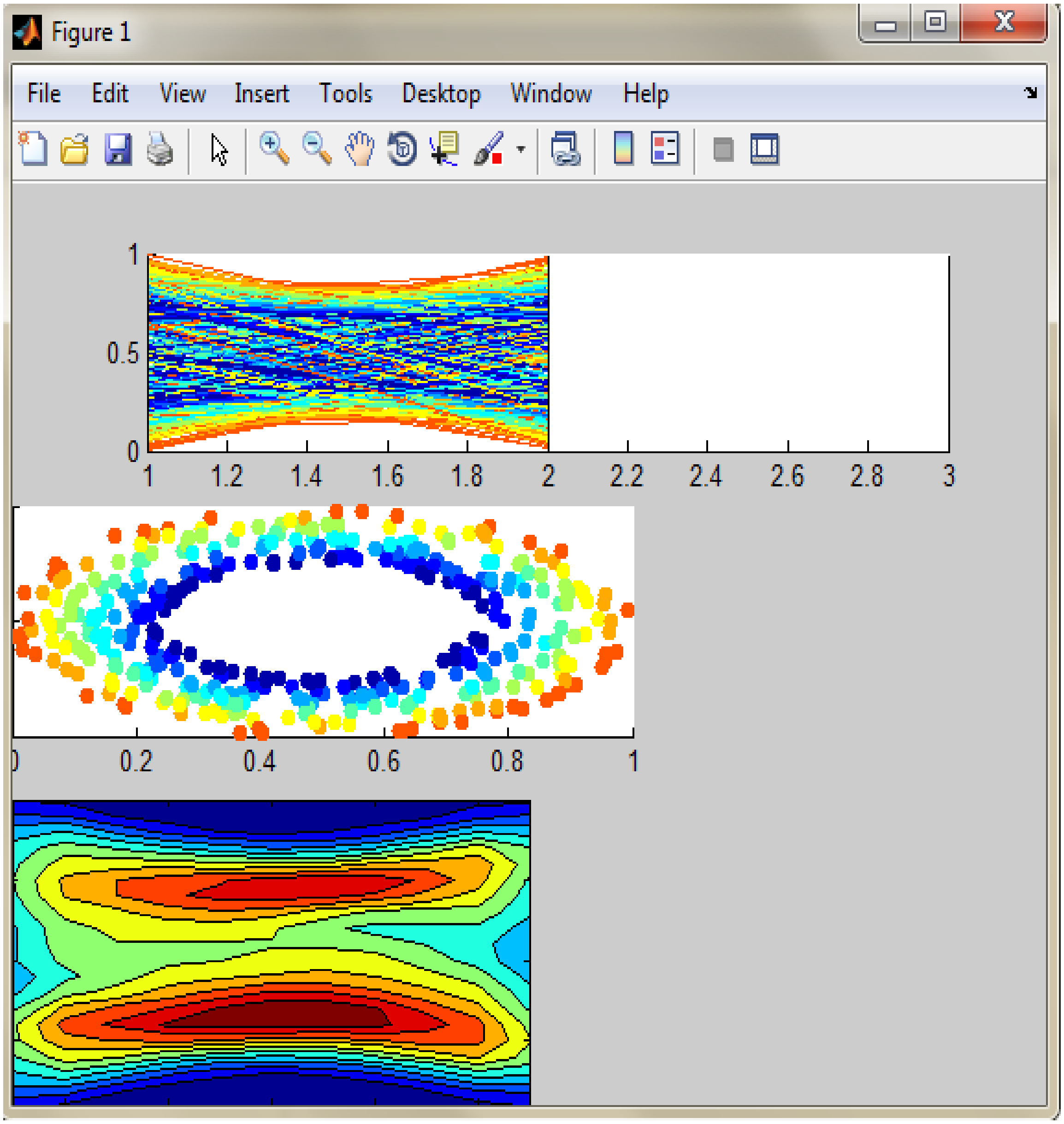
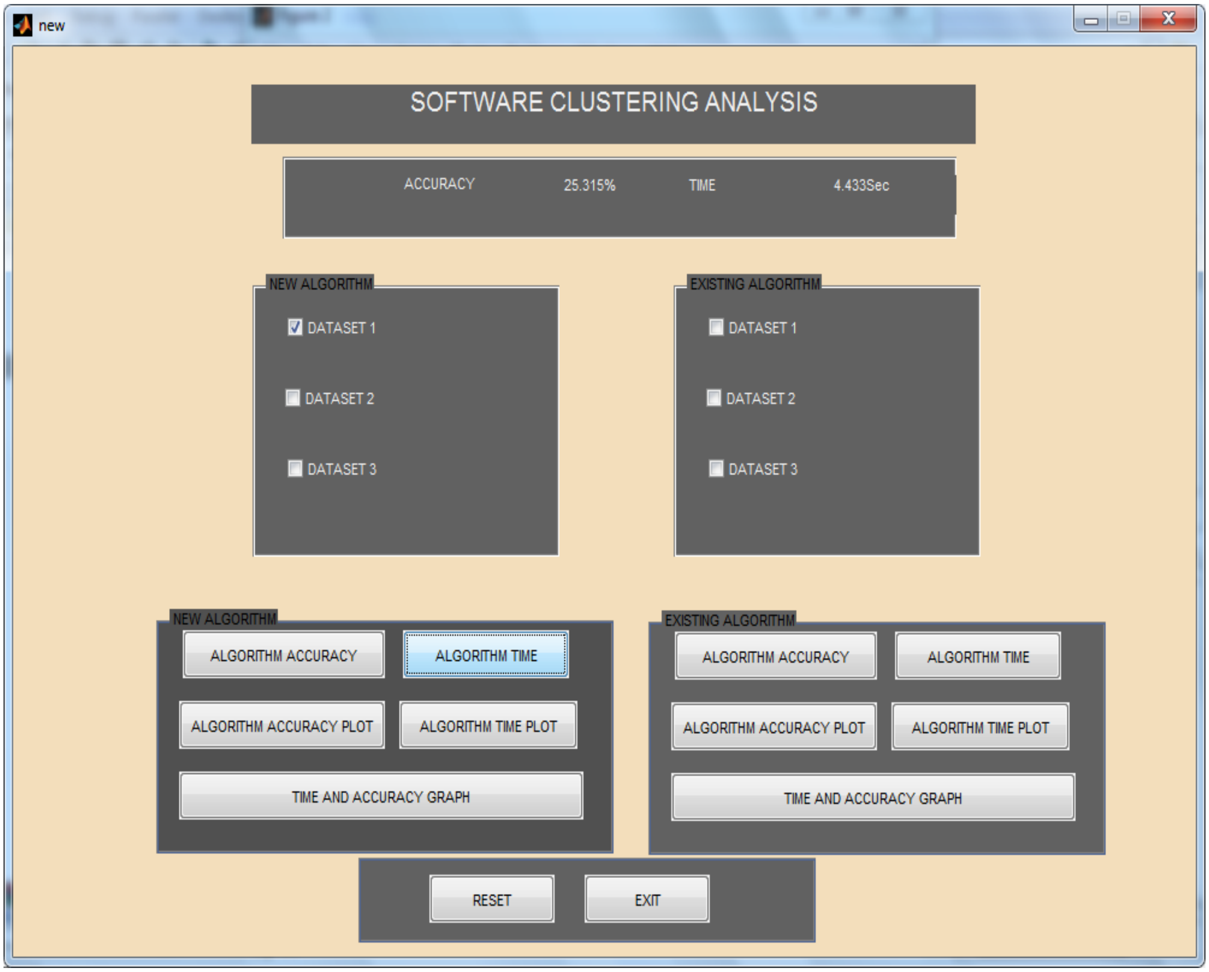


Fig. 6.10 Result of existing algorithm

As shown in figure 10, interface is designed for the performance analysis of existing k-mean algorithm. In existing algorithm data points are not clustered accurately



ACCURACY 25.315% TIME 4.433Sec

☑ DATASET 1
□ DATASET 2
□ DATASET 3

□ DATASET 1
□ DATASET 2
□ DATASET 3

Fig. 6.11 Performance analysis

As shown in figure11, interface is designed for the performance analysis of existing algorithm of k means and enhanced algorithm of k means. The performance analysis will be done in terms of accuracy and time of enhanced K-mean algorithm. In this figure accuracy on Dataset1 is 25.315 % and time is 4.433 sec

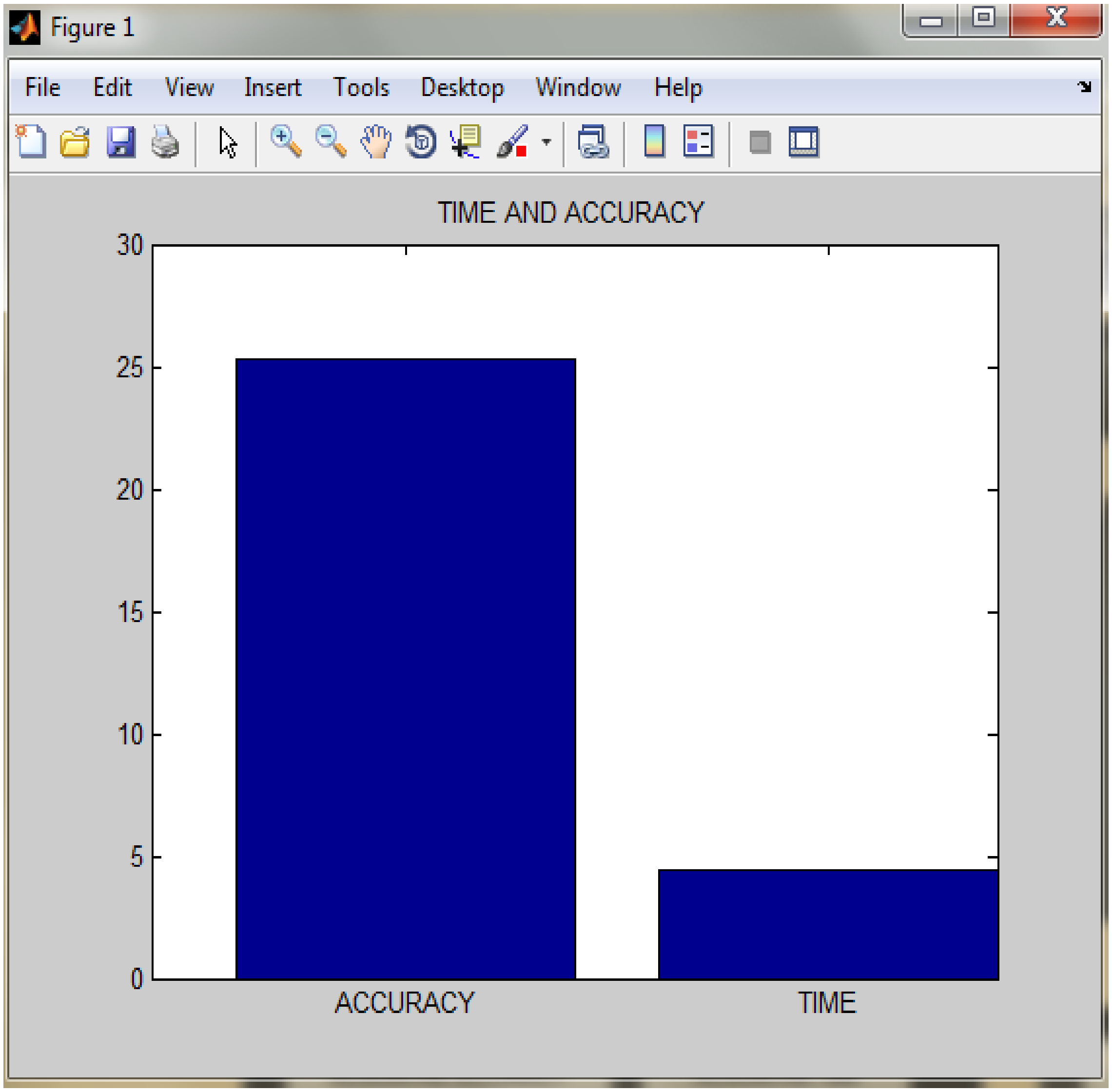
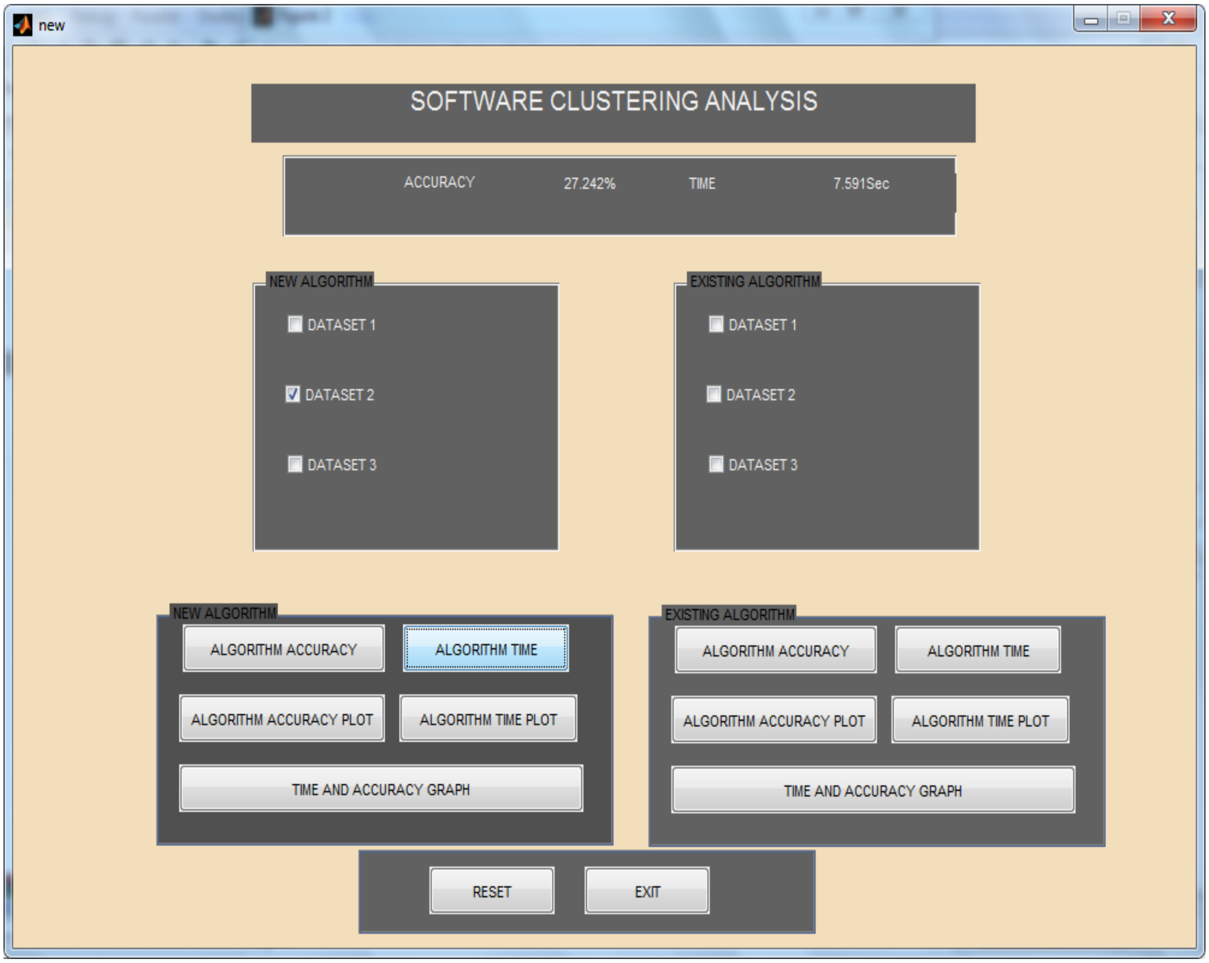


Fig. 6.12 Performance of enhanced algorithm

As illustrated in figure 12, the enhanced k-mean algorithm is performance show in form of bar graph. The accuracy on dataset 1 is 25.315 is % and time is 4.433 sec



SOFTWARE CLUSTERING ANALYSIS

ACCURACY 27.242% TIME 7.591Sec

NEW ALGORITHM

□ DATASET 1

☑ DATASET 2

□ DATASET 3

EXISTING ALGORITHM

□ DATASET 1

□ DATASET 2

□ DATASET 3

Fig. 6.13 Performance analysis

As shown in figure 13, interface is designed for the performance analysis of existing algorithm of k means and enhanced algorithm of k means. The performance analysis will be done in terms of accuracy and time of enhanced K-mean algorithm. In this figure accuracy on Dataset2 is 27.315 % and time is 7.433 sec

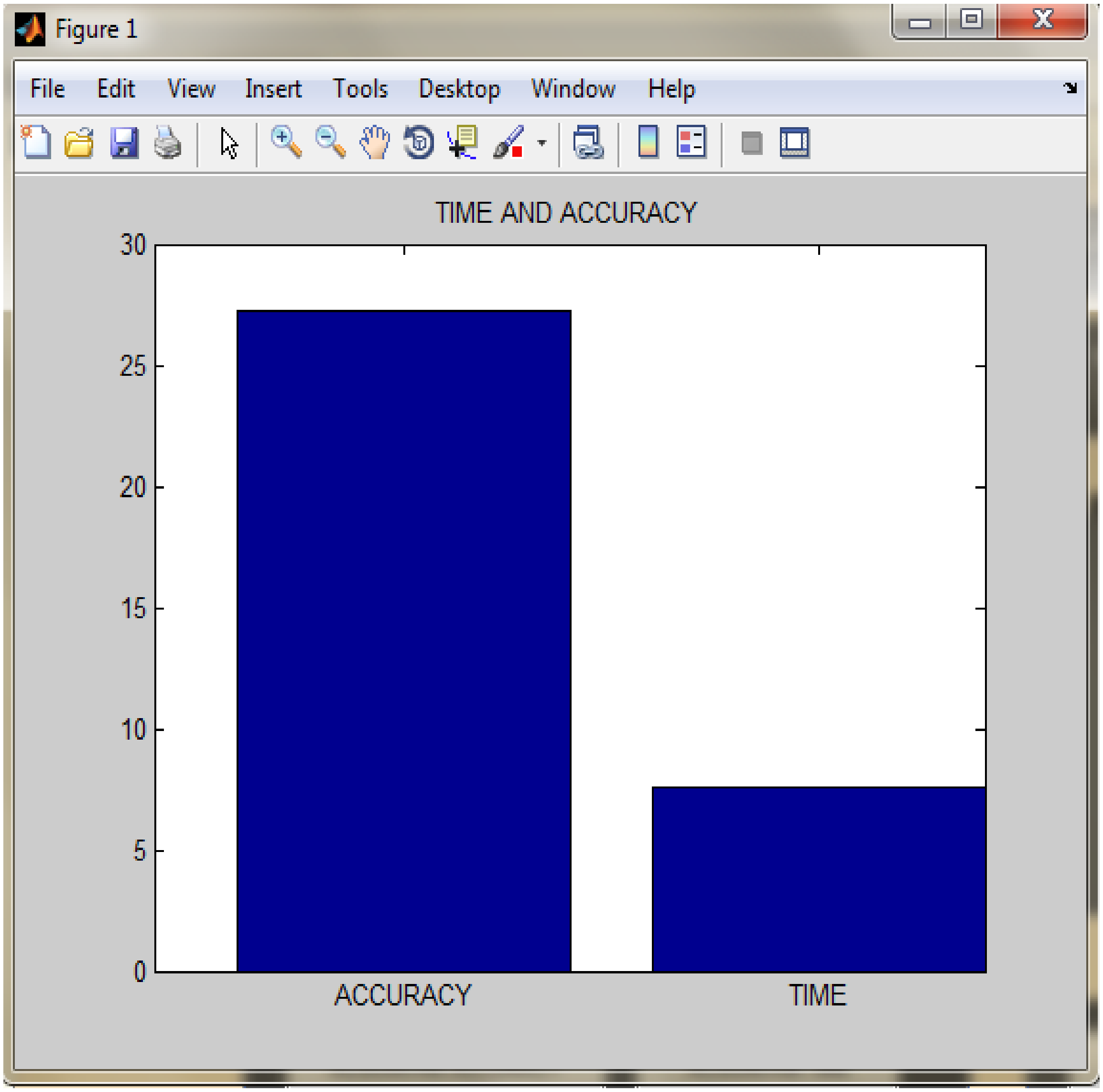
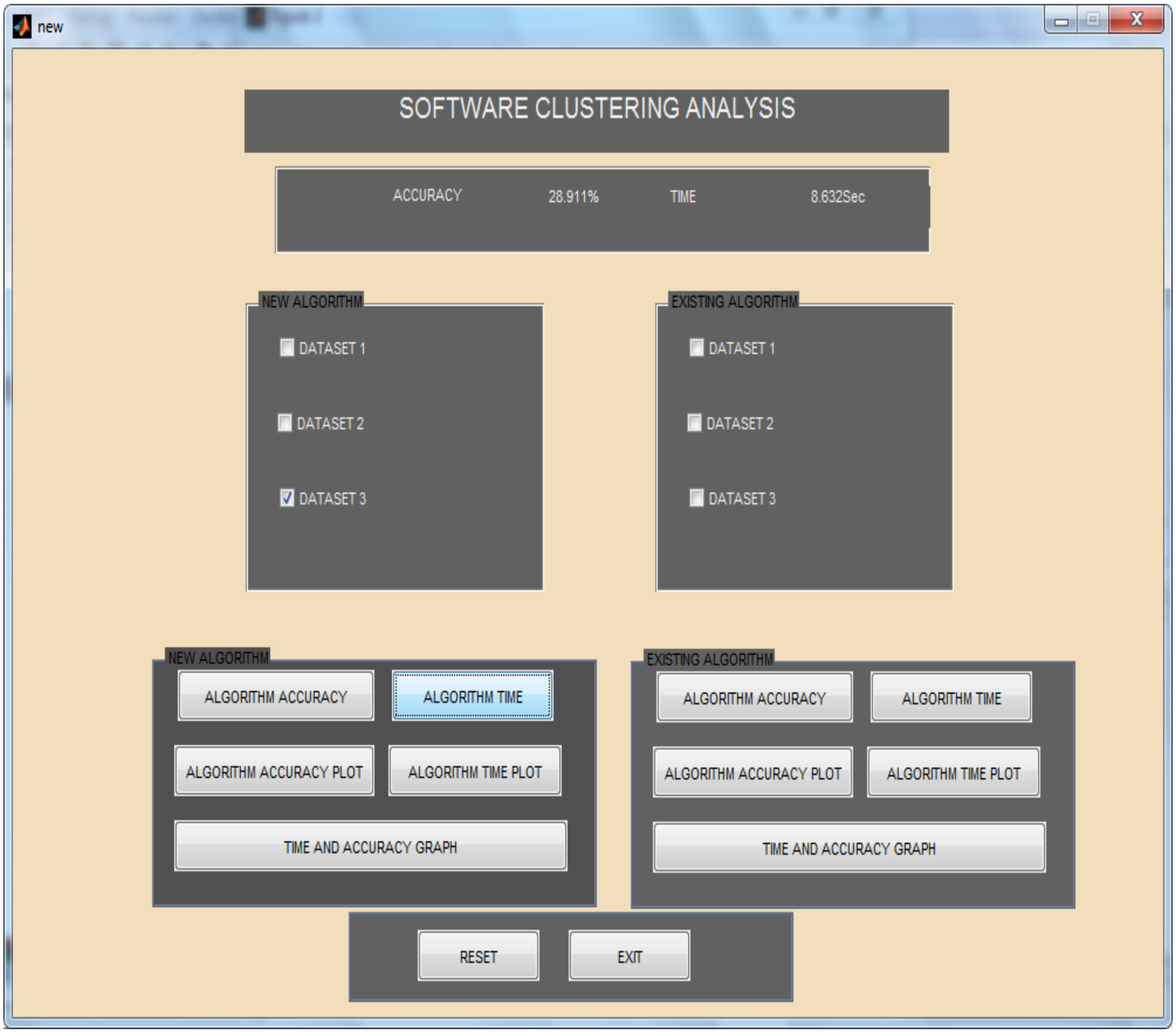


Fig. 6.14 Performance of enhanced algorithm

As illustrated in figure 14, the enhanced k-mean algorithm is performance show in form of bar graph. The accuracy on dataset2 is 27.315 is % and time is 7.433 sec



ACCURACY 28.911% TIME 8.632Sec

NEW ALGORITHM

□ DATASET 1

□ DATASET 2

☑ DATASET 3

EXISTING ALGORITHM

□ DATASET 1

□ DATASET 2

□ DATASET 3

Fig. 6.15 Performance analysis

As shown in figure 15, interface is designed for the performance analysis of existing algorithm of k means and enhanced algorithm of k means. The performance analysis will be done in terms of accuracy and time of enhanced K-mean algorithm. In this figure accuracy on Dataset3 is 28.911 % and time is 8.633 sec

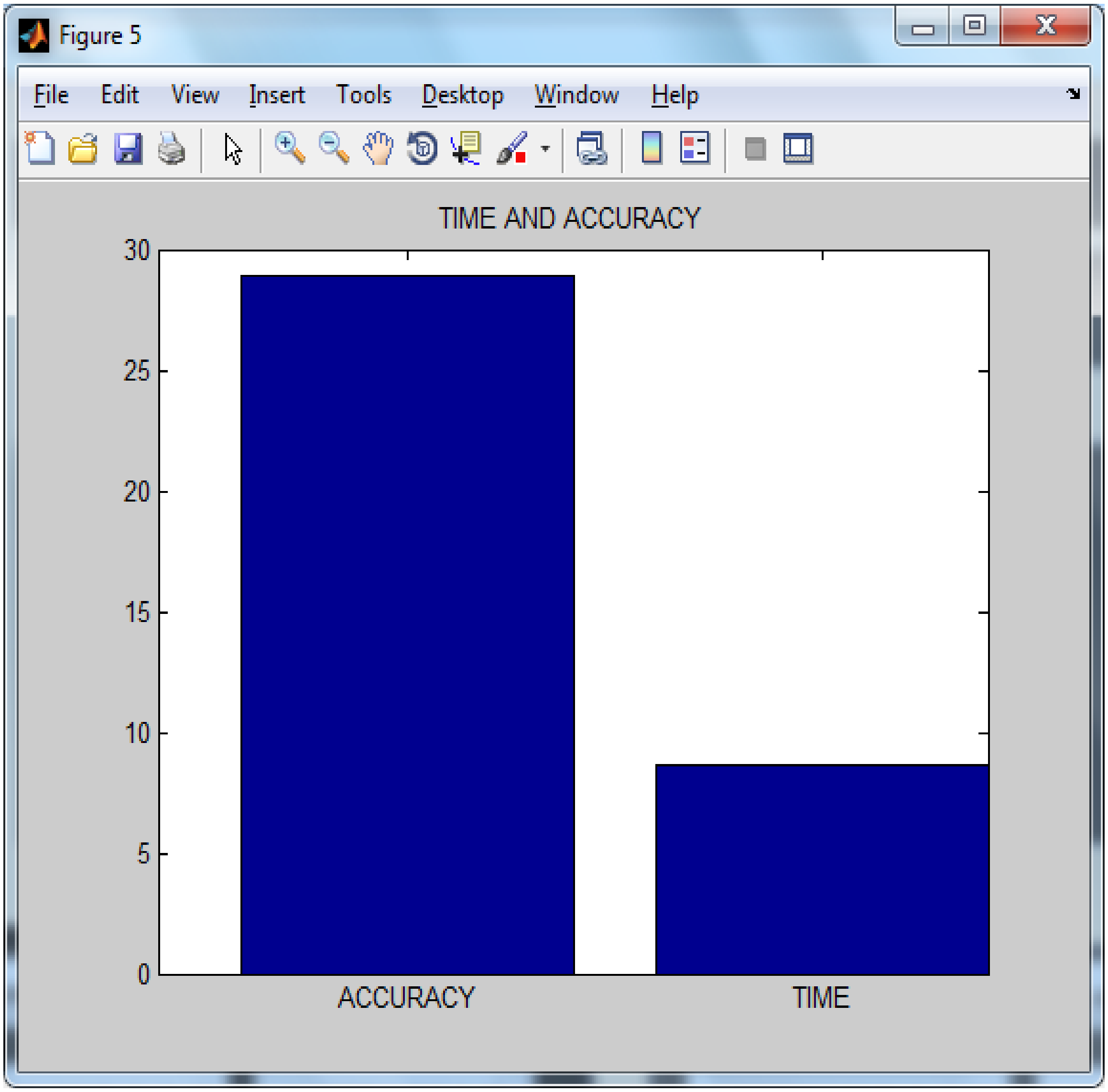


Fig. 6.16 Performance of enhanced algorithm

As illustrated in figure 16, the enhanced k-mean algorithm is implemented and performance show in form of bar graph. The accuracy on dataset3 is 28.911 % and time is 8.633 sec

Chapter-7

CONCLUSION

It is imperative to make technology decisions at the good time with good techniques and for the good logics. For Batter business suggests good people with proper supporting tools so they can develop very effective products. When it comes to establish the software, that time handle difficult language problem head-on is one constraint for today's creative manager. When combined with alternative software engineering applications, an effective language decision can support the cost-effective software systems development that, in turn, it arrange beneficial and effective, good support of business. In this paper presented the state of development and the evaluation methodologies of software clustering. We also describe the most valuable research challenges for this valuable research area. It should be also feasible that while the most valuable advances have already taken place, there are still many different paths for more research which will more effective for software engineers far and wide.

Research Paper

[1] Lingming Zhang, Ji Zhou, Dan Hao ,Lu Zhang, Hong Mei” *Prioritizing JUnit Test Cases in Absence of Coverage Information*” IEEE 2009.

[2] Paolo Tonella, Paolo Avesani, Angelo Susi” *Using the Case-Based Ranking Methodology for Test Case Prioritization*”. 22nd IEEE International Conference on Software Maintenance (ICSM'06),2009.

[3] Zheng Li, Mark Harman, and Robert M. Hierons” *Search Algorithms for Regression Test Case Prioritization*” IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 33, NO. 4, APRIL 2007.

[4] Praveen Ranjan Srivastava” *Test Case Prioritization*” Journal of Theoretical and Applied Information Technology,2008.

[5] Ruchika Malhotra, Arvinder Kaur and Yogesh Singh(June 2010)” *A Regression Test Selection and PrioritizationTechnique*” Journal of Information Processing Systems, Vol.6,No.2, 2010

[6] G. Pour, “*Component-Based Software Development Approach: New Opportunities and Challenges*,” Proceedings Technology of Object-Oriented Languages, 1998. TOOLS 26., pp. 375-383.

[7] Hans van Vliet, “*Some Myths of Software Engineering Education*”, Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands, 2010

[8] Joy B., Steele G., Gosling J., and Brach G., The Java Language Specification (2nd edition),ISBN 0-201-31008-2, Addison-Wesley, 2000.

[9] Shaheda Akthar1 , Sk.Md.Rafi , “*Improving the Software Architecture through Fuzzy Clustering Technique*” Vol 1 No 1 54-57

[10] Chih-Cheng Hung!, Wenping Liu and Bor-Chen Kuo, “*A new Adaptive fuzzy Clustering algorithm for remotely sensed images*” Marietta, GA 30060 USA

[11] WANG Jing1, TANG Jilong, “*Alternative Fuzzy cluster segmentation of remote sensing images based on adaptive genetic algorithm*” Chin. Geogra. Sci. 2009

[12] Markus Bauer Forschungszentrum Informatik Karlsruhe, "*Architecture-Aware Adaptive Clustering of OO Systems*" 2004 IEEE

[13] Zhang Chen et.al, "A *Robust Fuzzy Kernel Clustering Algorithm*" 1005-1012 (2013)

[14] D. Doval, S. Mancoridis, B. S. Mitchell , "*Automatic Clustering of Software Systems using a Genetic Algorithm*" Dept. of Mathematics and Computer Science 1999

[15] Narayan Desai, Rick Bradshaw, Ewing Lusk, "*Component-Based Cluster Systems Software Architecture a Case Study*" *ieeexplore.ieee.org* by N Desai - 2004

[16] Chung-Horng Lung, "*Software Architecture Recovery and Restructuring through Clustering Techniques*" by CH Lung - 1998

[17] Ioana S¸ ora, Gabriel Glodean, Mihai Gligor," *Software Architecture Reconstruction an Approach Based on Combining Graph Clustering and Partitioning*" IEEE by I Şora - 2010

[18] Kamran Sartipi ,"*Software Architecture Recovery based on Pattern Matching*" ON. N2L 3G1, Canada ksartipi

[19] Maninderjit Kaur and Sushil Kumar garg, *"Survey on Clustering Techniques in Data Mining for Software Engineering"*, International Journal of Advanced and Innovative Research (2278-7844) / # 238 / Volume 3 Issue 4 2014

[20 Manpreet Kaur and Usvir Kaur, "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", I*nternational Journal of Advanced Research in Computer Science and Social* , Volume 3, Issue 7, July 2013 ISSN: 2277 128X

[21] Tapas Kanungo , David M. Mount , Nathan S. Netanyahu Christine, D. Piatko , Ruth Silverman and Angela Y. Wu, "An Efficient K-Means Clustering Algorithm: Analysis and Implementation ," *IEEE Transactions on Pattern Analysis and Machine Intelligence,*

Volume 24, July 2002.

[22] Amar Singh and Navot Kaur, "To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm," *International journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 8, August 2012.

[23] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering , Vol IWCE 2009, July 1 - 3, 2009, London, U.K

[24] Shuigeng Zhou Yue Zhao Jihong Guan and Joshua Huang, *"NBC: A Neighborhood-Based Clustering Algorithm",* 2006

[25] Neha Aggarwal, Kirti Aggarwal and Kanika Gupta, "Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining," *International Journal of Scientific & Engineering Research*, Volume 3, Issue 3, August-2012.

[26] Yugal Kumar and G.Sahoo, " *A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm",* International Journal of Advanced Science and Technology Vol.62, (2014), pp.43-54, 2014.

[27] Jiawei Han, Micheline Kamber, Jian Pei, "Data minig concepts and techniques", Thrid edition.

[28] L.V.Bijuraj "clustering and its applications", Proceedings of National Conference on New Horizons in IT – NCNHIT, 2013