Enhancement in K-mean Clustering to Analyze Software Architecture Using Normalization

Preeti Puri*, Isha Sharma**

* [1]Post Graduate Student, Dept. of Computer Science Engineering (Software Engineering), Chandigarh University, Mohali, Punjab, 140413
Er.preeti1991@gmail.com
** Assistant Professor, Dept. of Computer Science Engineering, Chandigarh University, Mohali, Punjab, 140413
Ishasharma211@gmail.com

Abstract: Software engineering deals with the all kind of software production, design to coding, software accuracy and deals with the complexity of any software system. 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.

Keywords: K-mean clustering, Genetic algorithm, centre based clustering, efficiency, accuracy

I. Introduction

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 [1]. 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 [4]. Even big industries spend large some of amount on their software engineer for the software development [1].

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.

1.1. Clustering in Software Engineering: The basic idea behind clustering is to create the groups together for similar 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 [7]. 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.

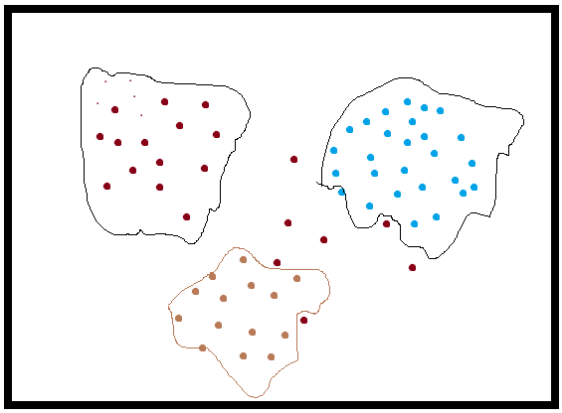

Fig 1.1: Clustering in Software Engineering

1.1.2 Types of Clustering: There are different types of clustering which are described as follows:

1.1.2.1 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. In well shaped cluster the data nodes are closer to each other. With this we can identify the data nodes and the values can be extracted. Sometimes threshold can be used to specify similarity or looseness between the nodes in cluster[9].

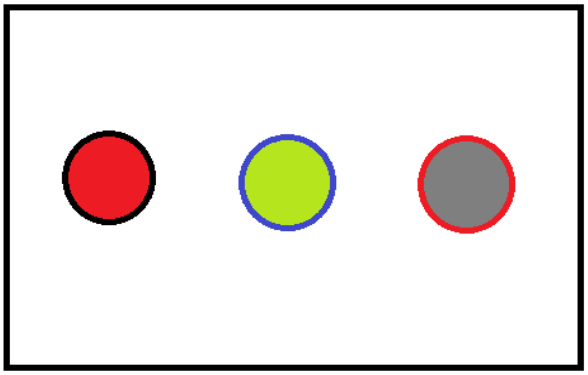

Fig.1.2 Well shaped cluster

1.1.2.2 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 center 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 center of a cluster [9].

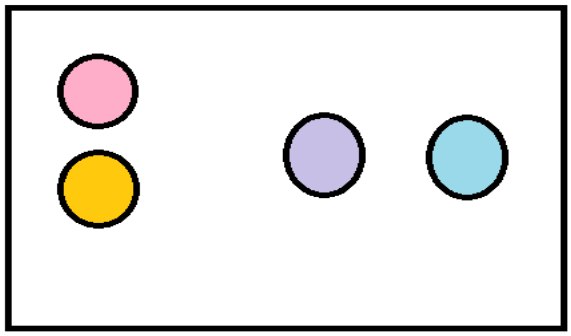

Fig.1.3: Centre based Cluster

3. Partitioning Methods: The general criterion for partitioning is a combination of high similarity of the samples inside of clusters with high dissimilarity between distinct clusters. Most partitioning methods are distance-based. Given k, the number of partitions to construct, a partitioning method creates an initial partitioning and then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another [10]. In a good partitioning the objects in the same cluster are close or related to each other whereas objects in different clusters are far apart or different. Most applications adopt popular heuristic methods such as greedy approaches like the k-means and k-medoids algorithms which progressively improve the clustering quality and approach a local optimum.

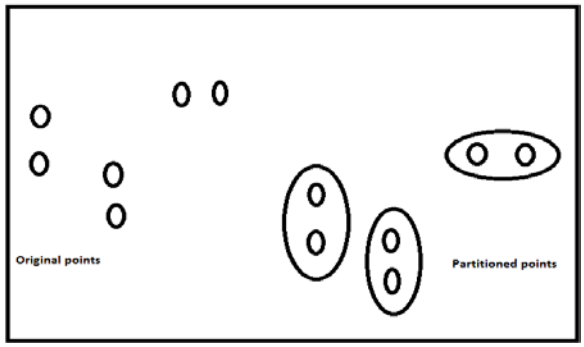

Fig.1.4 Partitioning Clustering

These clustering methods works well for finding spherical –shaped clusters in small to medium size

databases [8]. In this construct a partition of a data set containing n objects into a set of k clusters, so to minimize a criterion θ.The goal is, given a k, find a partition of k clusters that optimizes the chosen partitioning criterion. Here k is a input parameters. E.g. K-mean and K-centroid [10].

II. Review of Literature

In this paper [1] 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 the 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.

In this paper [2] 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.

In this paper [3] 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.

In this paper [4] 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.

In this paper [5] work represents ranking based method that improved K-means clustering algorithm performance and accuracy. In this they have also done analysis of K-means clustering algorithm, one is the existing K-means clustering approach which is incorporated with some threshold value and second one is ranking method which is weighted page ranking applied on K-means algorithm, in weighted page rank algorithm mainly in links and out links are used and also compared the performance in terms of execution time of clustering. Proposed ranking based K-means algorithm produces better

results than that of the existing k-means algorithm.

III. K-Mean Clustering:

The k-means clustering algorithm is the basic algorithm which is based on partitioning method which is used for many clustering tasks especially with low dimension datasets. It uses k as a parameter, divide n objects into k clusters so that the objects in the same cluster are similar to each other but dissimilar to other objects in other clusters. The algorithm attempts to find the cluster centres, (C1 …… Ck), such that the sum of the squared distances of each data point, xi , $1 \le i \le n$, to its nearest cluster centre Cj, $1 \le j \le k$, is minimized. First, the algorithm randomly selects the k objects, each of which initially represents a cluster mean or centre. Then, each object xi in the data set is assigned to the nearest cluster centre i.e. to the most similar centre. The algorithm then computes the new mean for each cluster and reassigns each object to the nearest new centre [24]. This process iterates until no changes occur to the assignment of objects. The convergence results in minimizing the sum-of-squares error that is defined as the summation of the squared distances from each object to its cluster centre [3, 9]. The following procedure summarizes the k-means algorithms [12]:

Algorithm: k-means:-The k-means algorithm for partitioning, where each cluster's centre is represented by the mean value of the objects in the cluster.

Input:

k: the number of clusters,

D: a data set containing n objects.

Output:

A set of k clusters.

Method:

(1) Randomly choose k objects from D as the initial cluster centres;

(2) repeat

(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;

(5) until no change;

IV. Proposed Methodology

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 previous work, 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.

V. Experimental Results

K-Mean clustering is the basic algorithm which is used to cluster items and work efficiently. In this work, MATLAB has used for implementation.

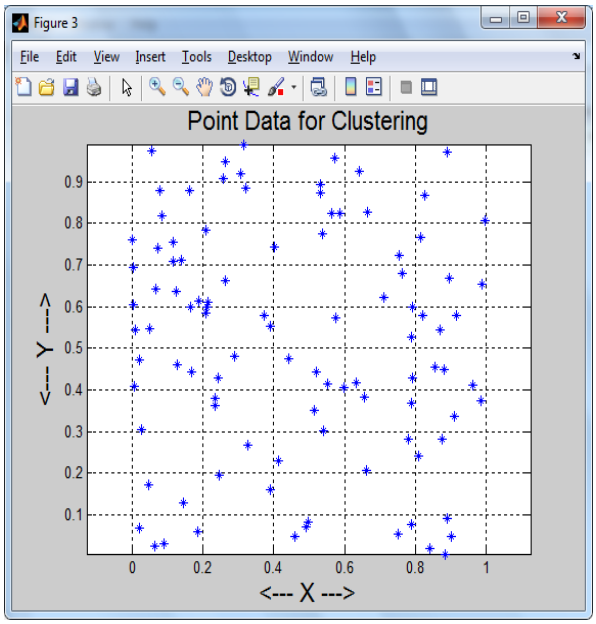

Fig 1.5: K-mean clustering

As shown in figure 1.5, 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.

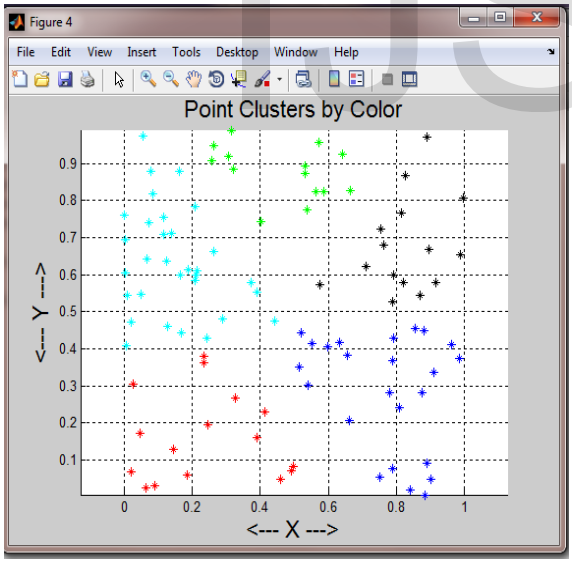

Fig 1.6: colored division clustering

As shown in figure 1.6, the formula of Euclidian distance was applied to cluster similar points in the k- mean clustering. After that the color division of the data is done so that the data can be easily identified.

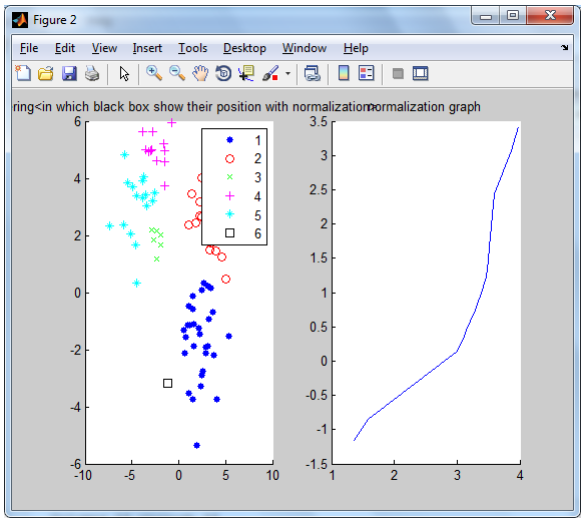

Fig.1.7 Enhanced K-mean clustering algorithm

As shown in figure 1.7, 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

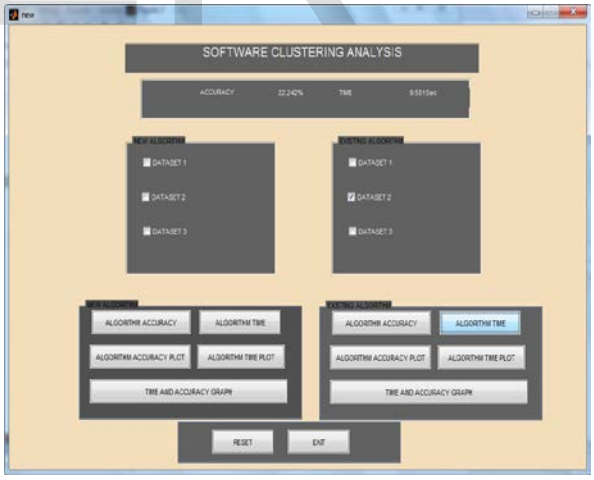

Fig.1.8 Performance analysis

As shown in figure 1.8, interface is designed for the performance analysis of existing k-mean algorithm and enhanced k-mean algorithm. 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

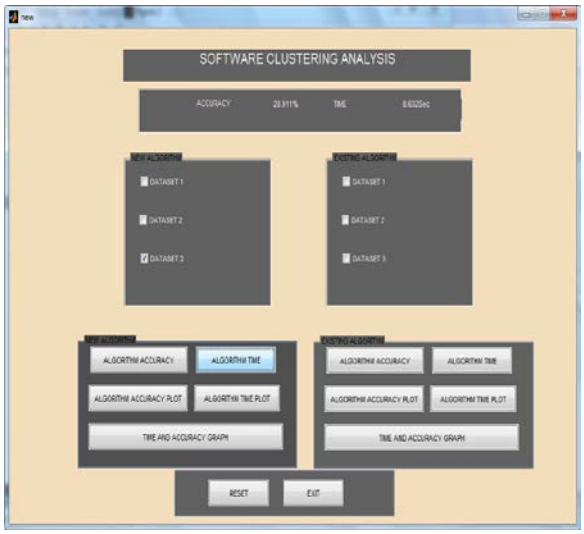

Fig.1.9: Performance analysis

As shown in figure 1.9, interface is designed for the performance analysis of existing k-mean algorithm and enhanced k-mean algorithm. 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.

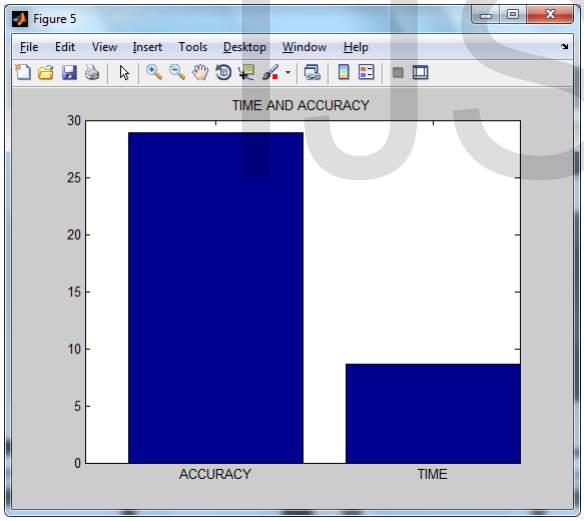

Fig 1.10: Performance of enhanced algorithm

As illustrated in figure 1.10, 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.

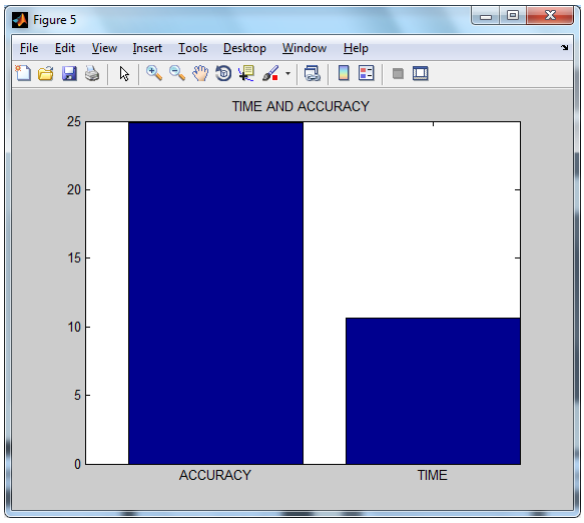

Fig.1.11 Performance of k-mean algorithm

As illustrated in figure 11, the enhanced k-mean algorithm is implemented and performance show in form of bar graph. The accuracy on dataset3 is 25.911 % and time is 13.633 sec

VI. 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. In previous work genetic algorithm had been used which is less efficient than K-mean. In this paper 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.

VII. FUTURE WORK:

In our work we hope that researchers do certain experiments in this field and discover some superior methods which results more efficient than this. There are many learning methods which can

produce better results. Data mining is a dynamic field on which the data miners are working and this field is continuously developed.
Data flow diagrams in this can help in easy differentiation of the data.

References

Lingming Zhang, Ji Zhou, Dan Hao ,Lu Zhang, Hong Mei" Prioritizing JUnit Test Cases in Absence of Coverage Information" IEEE 2009.

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.

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] 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.

[5] 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

G. Pour, "Component-Based Software Development Approach: New Opportunities and Challenges," Proceedings Technology of Object-Oriented Languages, 1998. TOOLS 26. pp. 375-383.

Hans van Vliet, "Some Myths of Software Engineering Education", Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands, 2010

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

Shaheda Akthar1 , Sk.Md.Rafi , "Improving the Software Architecture through Fuzzy Clustering Technique" Vol 1 No 1 54-57

Chih-Cheng Hung!, Wenping Liu and Bor-Chen Kuo, "A new Adaptive fuzzy Clustering algorithm for remotely sensed images" Marietta, GA 30060 USA

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 Informatics Karlsruhe, "Architecture-Aware Adaptive Clustering of OO Systems" 2004 IEEE

[13] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya," Comparison the various clustering algorithms of weka tools"," International Journal of Emerging Technology and Advanced Engineering" (ISSN 2250-2459, Volume 2, Issue 5, May 2012)

[14]Tajunisha and Saravanan, "Performance analysis of k-means with different initialization methods for high dimensional datasets, "International Journal of Artificial Intelligence & Applications (IJAIA), vol. 1, no.4, pp.44-52, Oct. 2010.

[15] D.Napoleon, S.Pavalakodi,"A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set," International Journal of Computer Applications (0975– 8887), vol. 13, no.7, pp.41-46, Jan 2011.

[16] Kehar Singh, Dimple Malik and Naveen Sharma, "Evolving limitations in K-means algorithm in data mining and their removal,"IJCEM International Journal of Computational Engineering &Management, vol. 12, pp.105-109, Apr. 2011.

[17] Dimitrios CharalampidisI,"A Modified K-Means Algorithm for Circular Invariant Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp.1856-1865, Dec 2005.

[18] Malay K. Pakhira,"A Modified k-means Algorithm to Avoid Empty Clusters," International Journal of Recent Trends in Engineering, vol. 1, no. 1, pp.220-226, May 2009.

[19] TapasKanungo, David M.Mount, Nathans. Netanyahu, Christine D.Piatko, Ruth Silverman, and Angela Y.Wu,"An Efficient-Mean Clustering Algorithm: Analysis an Implementation,"IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-891, Jul 2002.

[20] Fasahat Ullah Siddiqui and Nor Ashidi Mat Isa, "Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation,"IEEE, pp.833-841.

[21] Data mining and knowledge discovery handbook, Lior Rokach (Department of Industrial Engineering) Tel-Aviv University-Chapter15, CLUSTERING METHODS

[22] Soft Computing, A Fusion of Foundations, Methodologies and Applications

[23] Amerada Chug1, Shafali Dhall,"Software Defect Prediction Using Supervised Learning Algorithm and Unsupervised Learning Algorithm" (USICT, GGSIPU, New Delhi)

[24]N.S.Chandolikar,D.Nandavadekar,"Comparati ve Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Dataset, "International Journal of Computer Science and Engineering(IJCSE),vol.1,pp.81-88,Aug 2012.

[25] Mehmet Koyuturk, Ananth Grama and Naren Ramakrishnan,"Compression, Clustering and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets,"IEEE Transactions on Knowledge and A Data Engineering", vol. 17, no. 4, pp.447-461, Apr 2005.

[26] Christopher M. Bishop and Michael E. Tipping, "A Hierarchical Latent Variable Model For Data Visualization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp.281-293, Mar. 1998.

[27]Mu-Chun Su and Chien-Hsing Chou, "A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp.674-680, Jun. 2001.