ISSN:0975-9646

To Enhance A-KNN Clustering Algorithm for Improving Software Architecture

Ishu
Student, Lovely Professional University

Sandeep Singh
Assistant Professor, Lovely Professional University

Abstract: Software Architecture is important factor for the development of complex and big software system. Software Architecture Decomposition is an important part in software design. Software clustering is used to cluster functions of similar type in one cluster and other are in other cluster. K-mean is the base of the clustering but it has some limitations. Many clustering methods are used for decomposition the software architecture. A-KNN cluster method is more efficient than others methods but some functions are highly coupled then cluster technique does not find out correct distance. So that need to enhancement in Euclidian distance formula based on normalization. In this paper, an enhancement has proposed in the Euclidean distance formula which has increased the cluster quality. When the cluster quality will be increase then cluster the highly coupled function properly and improve the software architecture and A-KNN will be generate the best results than previous method.

Keyword: **Architecture, A-KNN, Clustering, Decomposition.**

1. INTRODUCTION

Software is a not tangible device like computer programs and documentation. It is not similar from other tangible hardware device. Software Engineering is the discipline of computer science which follows some principles to create, change and maintain of software elements [1]. Software Engineering is a set of problem solving Skills, instructions and methods applied upon a variety of domains to discover and create useful systems that is used to solve practical problems [5]. Software engineering is series of steps to create the software, from its preliminary stage to its last stage. Software is a generic term that is used for organized the data and instructions that are collected to develop it. Software Engineering is a work to provide high quality Software products to its customers.

1.1 Software Architecture:

Software Architecture refers to the high level structures of a software system, the discipline of creating such structures, and the documentation of these structures [5]. Software architecture is one of the most important factors for the development of complex and big software systems. Architecture is a structure of the system which comprise software element, external feasible properties of those element and relationship among those component [6].

1. Architecture is the overall structure of the system. It is the structure of the component of a program or system.
3. Architecture is about fundamental thing.
4. Architecture is component and connector.

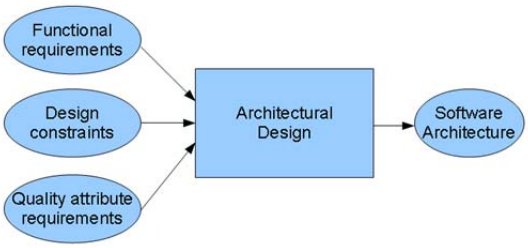


Figure 1. Software Architecture

1.2 Clustering in Software Engineering: Clustering is an unsupervised classification method aims at creating groups of objects, or clusters, in such a way that objects in the same cluster are very similar and objects in different clusters are quite distinct. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing [4]. In business, clustering can help marketers discover interests of their customers based on purchasing patterns and characterize groups of the customers. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. In geology, specialist can employ clustering to identify areas of similar lands , similar houses in a city and etc. data clustering can also be helpful in classifying documents on the Web for information discovery [7,8]. The method which is used to group of same type of documents is called clustering. There are some other benefit of clustering is that documents can show in multiple subtopics, thus we are sure that a useful and necessary document will not be misplaced from other search results. When a vector of topic for every document and analyze the weights of how fit the document into each cluster is a basic clustering algorithm.

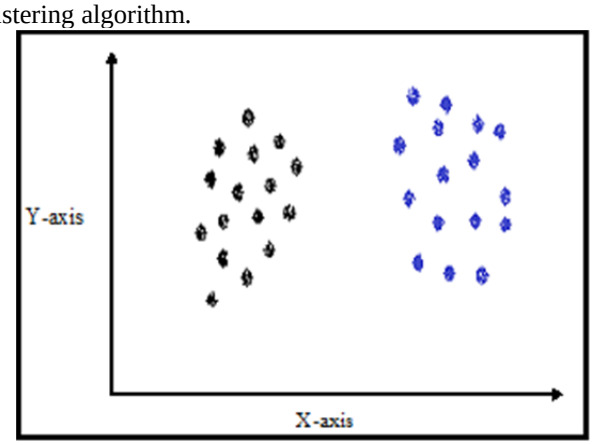


Figure 2. Clustering

1.2.1 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 [8]. 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 k-means algorithm has following drawbacks:

1. As many clustering methods, the k-means algorithm assumes that the number of clusters k in the database is known beforehand which, obviously, is not necessarily true in real-world applications.
2. As an iterative technique, the k-means algorithm is especially sensitive to initial centres selection.

2. REVIEW OF LITERATURE

In this paper [1] they explained about Software architecture decomposition plays an important role in software design. . Decomposition means that large or complex problem broken down into parts.. This paper presents clustering techniques for software architecture decomposition. There are uses two Hierarchical Agglomerative Clustering and adaptive K-nearest neighbour algorithm .It applied on two industrial software systems. There are uses two method to finding the distance between clusters SLINK and WPGMA. SLINK (Single Linkage algorithm) finds the distance between the nearby pair of components; take a one component from each cluster. WPGMA (Weighted pair group method using arithmetic averages) finds the distance among two clusters is taken as the average of distance between all pairs of components in the two clusters. In this approach, software architecture decomposition is done by clustering techniques with focus on functional requirements and attributes of the software. This paper presents an enhanced approach for decomposition of software architecture. It introduced the use of A-KNN algorithm in software architecture decomposition with focus on functional requirements and attributes and compared its performance with two Agglomerative Clustering algorithms or techniques: SLINK and WPGMA. The results demonstrate A-KNN algorithm is competitive with two Agglomerative Clustering techniques. It provides valuable information for software designer.

In this paper [2] they explained when apply the clustering techniques to software system decomposition, the software designer have two problems :(1) determination of number of clusters (2) determination of specific cluster or software module for some highly coupled or fuzzy component. This paper presents approach for finding solution to those two issues. There are used fuzzy C-mean clustering with three hierarchical agglomerative clustering techniques and the adaptive K- nearest neighbour algorithm. It applied on real industrial software systems. There are uses three method to finding the distance between clusters. SLINK, CLINK and WPGMA. SLINK (Single Linkage algorithm) finds the distance between the closest pair of components, taking one component from each cluster. WPGMA (Weighted pair group method using arithmetic averages) finds the distance between two clusters is taken as the average of distance between all pairs of components in the two clusters. CLINK (Complete linkage algorithm) finds the distance between most distant pair of component, taking one component from each cluster. Fuzzy C-mean clustering identify "fuzzy component" by membership clusters. . They introduced the use of A-KNN algorithm in software architecture decomposition using requirements and attributes and compared its performance with three agglomerative clustering algorithms: SLINK, CLINK and WPGMA. They conducted a set of experiments using two industrial software systems. Results of this approach shows A-KNN is competitive than others three agglomerative clustering techniques and FCM provide valuable information, which is helpful to handle the two issues for clustering techniques. In this paper [3] various software clustering algorithms is developed with its properties, qualities and restrictions. These algorithms used for specific software systems, but the main factor, how to choose a clustering algorithm which is best suitable for specification software. In this paper, it provides a way for the choosing of a software clustering technique for particular requirements. Software clustering is a domain that has been developing for a long time. Several software clustering algorithms have been proposed. In this paper, focus on software clustering techniques or methods that collect software elements into subparts. It is important think to understand the software system for someone. The outcome of a software clustering algorithm is known as decomposition of the software. The selection of a software clustering algorithm plays an important role for creating the decomposition. This paper provides the method for choose the appropriate algorithm. This approach allows the evaluation of the new algorithm in earlier stages before its implemented.

3. KNN AND A-KNN APPROACHES

3. 1 KNN Approach: It is an enhancement of K-mean clustering. It is based upon normalization.KNN is a non parametric lazy learning algorithm. It is very easy to understand but hard to implement. Non-parametric statement means that it does not make any assumptions on the underlying data distribution [7]. It makes decision on the basis of entire training data set. It has minimal training phase but a costly testing phase. Cost is in terms of memory and time. It requires more time to access all the data training sets. It also requires more memory to store all the data.

KNN assumes that data is in feature space and data points are metric points. The data can be scalars and multidimensional vectors. Each training set is consisting of vector and each vector has label like positive and negative. But KNN works equally with arbitrary numbers. It has value of K. The value of K decides the neighbours of the classification. If the value of K=1 then this algorithm is simply known as nearest neighbour algorithm. It is done to increase the quality of the clustering. It can be used to find out the density estimation of the classification. In Figure 3, first class is blue because of cluster having more blue points and other class is red because of red cluster having more red points than blue.

KNN approach is basically consisting of following steps:

1. Data set uploads.
2. Find out probabilistic points in which maximum points are lies near centre known as hyper plane.
3. Find out Euclidean distance from hyper plane.
4. Plotting the points on the basis of Euclidean distance.

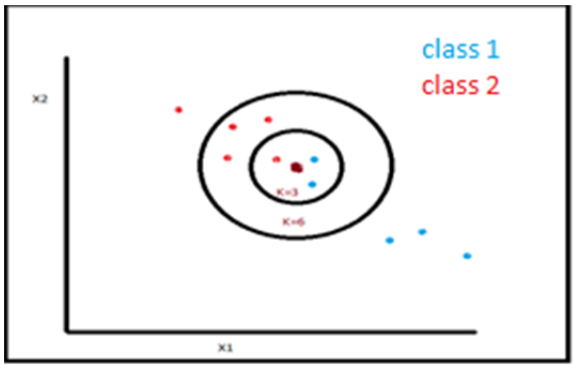


Figure 3.KNN Approach

Disadvantages of KNN: Its main disadvantages are as follow:

1. KNN needs to determine the value of parameter K.
2. It is not clear which distance based learning will produce best results. It is hard to decide which attribute contribute more and which contribute less.
3. Computation cost is very high due to the calculation of the each training set.

3.2 A-KNN Approach*:* Clustering is a technique which is used to assign the elements of similar properties in one cluster and cluster of different properties in another cluster. Clustering is a technique that is used to find out the elements in a data set efficiently. Clustering is an effective in multi dimensionally that is difficult to arrange in effective manner in other environment. The traditional technique of clustering is k-mean clustering. It has disadvantage that it is not easy to identify the initial of k seeds. The main advantage of A-KNN clustering over the HAC is the reduction the amount of the computations [1, 2]. A-KNN algorithms initially believe every entity as cluster. The labelled identity with a unique identifier always used to present the cluster identity. Second iteration there which assume that K=3, here K is the number of nearest neighbours (NN) to be choose by the user. The algorithm selects the K=3 nearest neighbours to the entity that will be clustered, and then checks labels. Out of three cluster when comes two clusters have the same label, the current entities having the same level in those two entities with the algorithm. , different label of those three entities are there then, the label of the algorithm of the current entity with the same label of the nearest label which is closest entity. When no more changes require in clustering tree then clustering process in algorithm repeats again. Output of the algorithm in one cluster is at the highest level of the hierarchy. The resemblance matrix in AKNN is intended only once. In adaptive K-mean, the number of elements in each cluster decides the number of comparison for each cluster [1, 2].

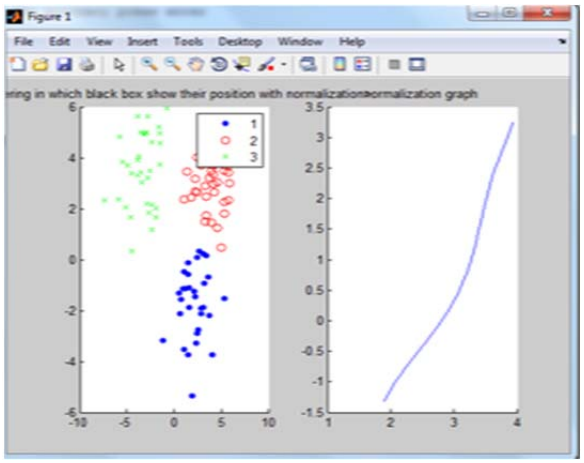
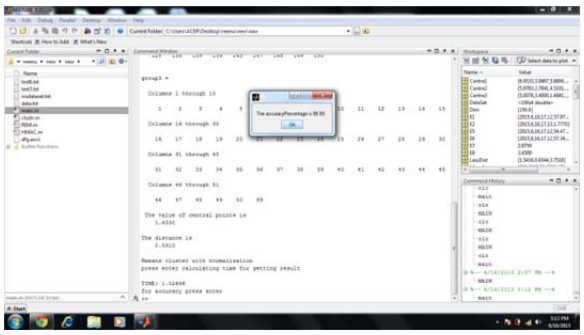


Figure 4 : Cluster the data by A-KNN



Fifure 5 : Accuracy of A-KNN

4. PROPOSED METHODOLOGY

A-KNN clustering algorithm is used to cluster the functions of the software for decomposition of software architecture. In A-KNN clustering algorithm, probability of the most relevant function is calculated and using Euclidean distance formula the functions are clustered. In this work, enhance in Euclidean distance formula to increase the cluster quality. The enhancement based on normalization. The proposed technique has been implemented in MATLAB. Firstly implement the A-KNN algorithm. Second is enhancement in Euclidean distance formula by using normalization technique. It makes A-KNN algorithm more accurate. A-KNN will properly cluster the highly coupled functions of software. Then implement the proposed technique and compared with previous A-KNN algorithm.

It is basic distance between two points in Euclidean space. The Euclidean distance between points P and Q is length of line segment connecting them.

When p = (p_1, p_2) and **q** = (q_1, q_2)

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

Algorithm

1. Input: Dataset of Software
2. Output: Clustered Data

[row column]=size(Dataset);

1. Load dataset and define number of iterations on the dataset
2. Assign number of clusters and assign members to each cluster on the basis of uniqueness

3. Define normalization point from the dataset using sigma function
4. Check number of members in each class
   If (Class1 members!=Class2 members)
Redefine the position of Normalization point
   Else
Assign final position of the normalization point
5. Apply Euclidean distance formula with normalization
6. Plot final results of data clustering

5. EXPERIMENTAL RESULTS

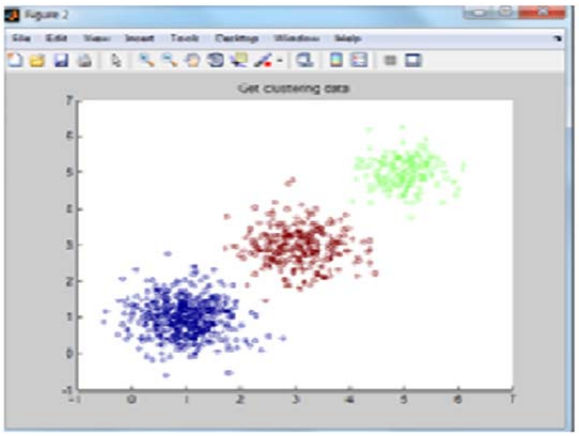


Figure 6: Plotting of data points with different colours

As illustrated in the figure 4, the A-KNN clustering algorithm is enhanced in which normalization is applied on the Software dataset to improve the cluster quality. In the figure clustered data is showed on 2D plane with three different colors.

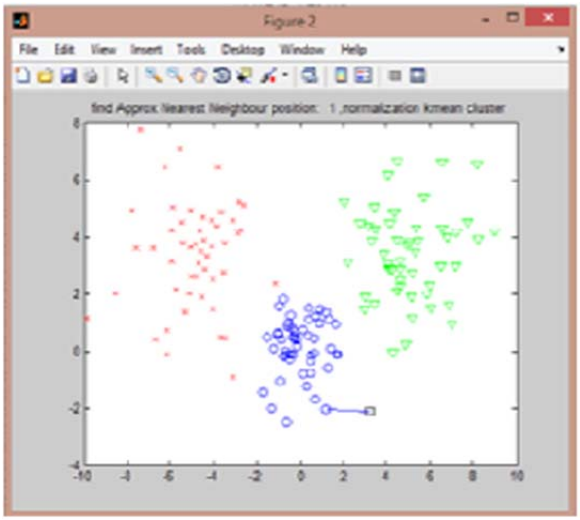


Figure 7: Normalization applied on Euclidean Distance

As illustrated in the figure 5, the A-KNN clustering algorithm is enhanced in which normalization is applied on the Software dataset to improve the cluster quality. In the previous figure clustered data is shown with three different colours and uniqueness of each cluster is calculated and data will be scattered according to their similarity on the 2D plane. From the scattered data one point is selected on the basis of probability which is marked with Black Square. From that square black point distance to each point in the cluster is calculated.

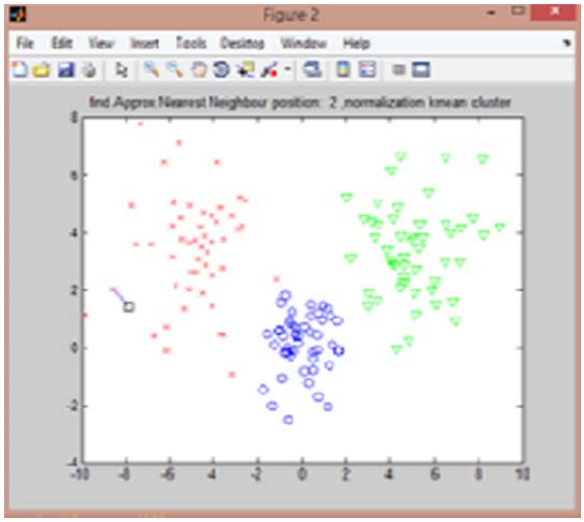


Figure 8. Distance from selected point to Data point

As illustrated in the figure 6, the A-KNN clustering algorithm is enhanced in which normalization is applied on the Software dataset to improve the cluster quality. In the previous figure clustered data is shown with three different colours and uniqueness of each cluster is calculated and data will be scattered according to their similarity on the 2D plane. From the scattered data one point is selected on the basis of probability which is marked with Black Square. From that square black point distance to each point in the cluster is calculated. The black point moved to another cluster and distance is calculated to another point.

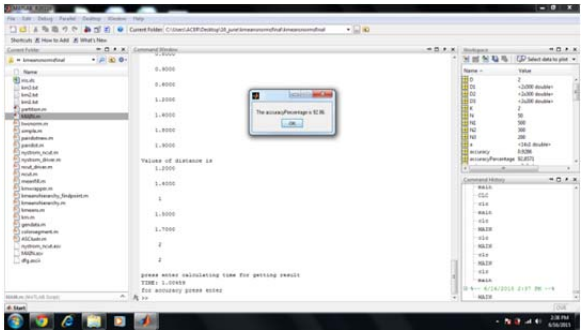


Figure 9. Accuracy of Enhanced A-KNN

As illustrated in the figure 7, the A-KNN clustering algorithm is enhanced in which normalization is applied on the Software dataset to improve the cluster quality. In the previous figure clustered data is shown with three different colours and uniqueness of each cluster is calculated and data will be scattered according to their similarity on the 2D plane. From the scattered data one point is selected on the basis of probability which is marked with Black Square. From that square black point distance to each point in the cluster is calculated. The black point moved to another cluster and distance is calculated to another point. Final position is calculated and on the basis of final point whole data is clustered. In the last accuracy of the final clustering is calculated.

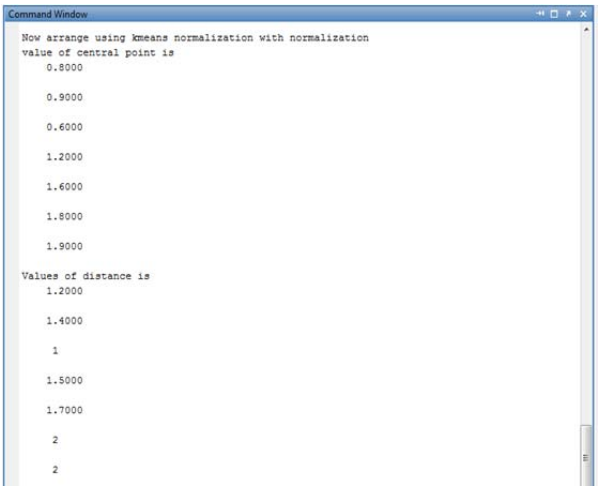


Figure:10,Values of centroid and Euclidean distance

As illustrated in the figure 8, After process of algorithm, find the values of centroid and distance between points.

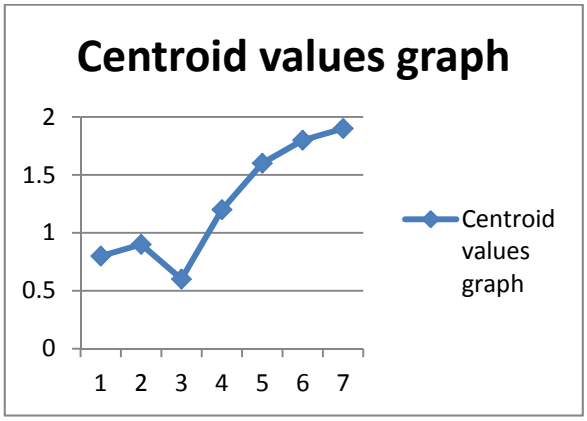


Figure 11: Centriod Values Graph

As shown in figure 9, the centroid point line graph is plotted at various iterations. The number of iteration is shown at x-axis and at y-axis centroid values are plotted .

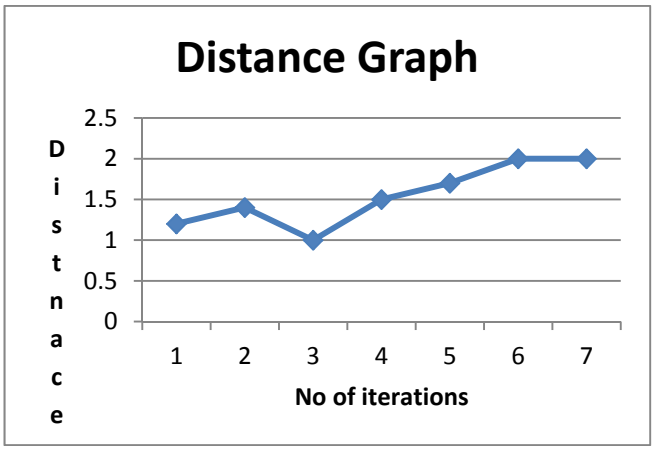


Figure 12: Distance Values Graph

As shown in figure 10, the Distance point line graph is plotted at various iterations. The number of iteration is shown at x-axis and at y-axis distance values are plotted .

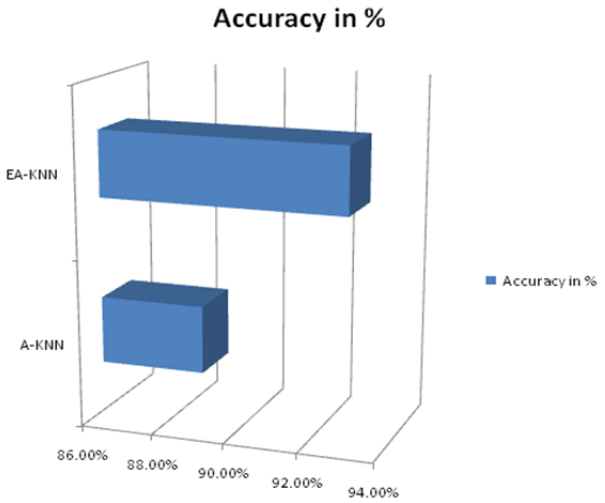


Figure 13, Comparison of Accuracy in graph

6. CONCLUSION

Software Engineering is sequence of steps to produce the software from initial stage to its final stage. Software architecture is important for develop the software. Software Architecture is most important factor for the development of complex and large software system. Architecture decomposition decreases the software complexity. Clustering is creating groups of objects, or clusters, in such a way that objects in the same cluster are very similar and objects in different clusters are quite distinct. Clustering results are helpful to architect or designer for decomposition of architecture.

Several clustering techniques are studied. A-KNN clustering technique gives the best result than others clustering technique. In this report, A-KNN algorithm has been used for decomposition of software Architecture with enhancing the Euclidean distance formula based on normalization that is increase the accuracy of A-KNN clustering technique. The proposed enhancement has been implemented for decomposition of software architecture focus on functional requirements and attributes. The results demonstrate that enhanced A-KNN is best then previous A-KNN clustering technique. Increase the clustering quality of A-KNN.

In Future work, Meta Clustering can be applying with normalization for improving the software quality. It will be increase the accuracy of algorithm. Clustering technique widely used to improve the software quality.

REFERENCES

[1] Abdulaziz Alkhalid, Chung-Horng Lung, Samuel Ajila, " Software Decomposition Using Adapative K-Nearest Neighbour Algorithm", 26th IEEE Canadian Conference Of Electrical And Computer Engineering (CCECE), 2013.

[2] Abdulaziz Alkhalid, Chung-Horng Lung, Duo Liu, Samuel Ajila, "Software Architecture Decomposition Using Clustering Techniques", IEEE 37th Annual Computer Software and Applications Conference, 2013.

[3] Mark Shtern and Vassilios Tzerpos, "Methods for Selecting and Improving Software Clustering Algorithms", 2014.

[4] Sadegh Bafandeh Imandoust And Mohammad Bolandraftar, **"**Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background", Int. Journal of Engineering Research and Application Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610.

[5] Timothy C. Lethbridge and Robert Laganiere, “Object-Oriented Software Engineering”.
[6] Len Bass, Paul clements and Rick Kazman, “Software Architecture in Practice”.
[7] L.V Bijuraj, “Clustering and its applications”, 2013.
[8] Neha Aggarwal, Kirti Aggarwal and Kanika Gupta, “Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining,” .
[9] R. Braden, L. Zhang, S. Berson, S. Herzog and S. Jamin, “Resource ReSerVation Protocol” (RSVP), 1997.
[10] Abdulaziz Alkhalid, Mohammad Alshayeb and Sabri A. Mahmoud, “Software Refactoring at the Class Level using Clustering Technique”, 2011.
[11] Abdulaziz Alkhalida, Mohammad Alshayebb, Sabri Mahmoudb, “Software refactoring at the function level using new Adaptive K-Nearest Neighbor algorithm”,2010.
[12] Chung-Horng Lung, Xia Xu, “Software Architecture Decomposition Using Attributes”, 2007.
[13] Chung-Horng Lung, Xia Xu, “Software Architecture Decomposition Using Attributes”, 2005.