ENHANCEMENT IN A-KNN CLUSTERING TECHNIQUE TO ANALYSE SOFTWARE ARCHITECTURE

Satwinder Singh,[1] Kamaljeet Kaur[2]
[1]Assistant Professor, [2]Research Scholer
[1,2]Department of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, (Punjab), India

***ABSTRACT*: Software Architecture is important factor for the development of complex and big software system. Software clustering is used to cluster functions of similar type in one cluster and others are in other cluster. K-mean is the base of the clustering but it has some limitations. A-KNN, WPGMA, UWPGMA, CLINK and SLINK these clustering methods are used for decomposition the software architecture. A-KNN cluster method is more efficient than other methods but some time 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 increased we are able to properly cluster the functions and improve the software architecture and A-KNN will be generating the best results than previous methods. It has covered the concept of cohesion and coupling between components of the system.**

***KEYWORDS*: K-mean, KNN, A-KNN, Clustering, Decomposition**

I. INTRODUCTION

Software is a not tangible device like computer programs and documentation. It is different from other tangible hardware device. Software Engineering is the discipline of computer science which follows engineering principles to create, operate, change and maintain of software components [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 [12]. Software engineering is all about sequence of steps to produce the software, from its initial stage to its final stage. A software engineering is related to all the aspects that are used in the software production or create the software. Software is a generic term that is used for organizing the data and instructions that are collected to develop it.

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 [11]. Software architecture is one of the most important success 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.

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.

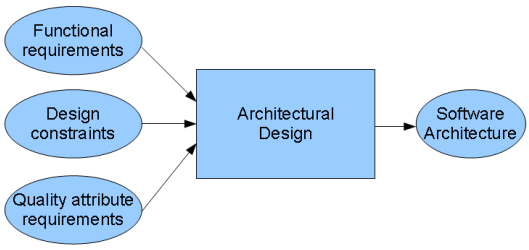

Fig.1. Software Architecture

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 [8]. Clustering is a method used to group similar documents, but it differs from categorization of documents are clustered on the fly instead of through the use of predefined topics. Another advantage of clustering is that

documents can appear in multiple subtopics, thus ensuring that a useful document will not be misplaced from search results. A basic clustering algorithm forms a vector of topics for each document and measures the weights of how healthy the document fits into each cluster.

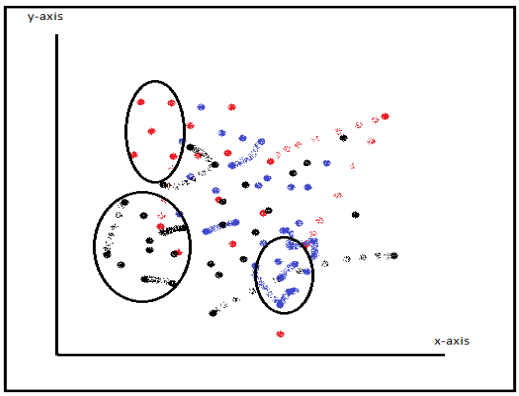

Fig.2 Clustering

1.3 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 [10]. 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. Despite being used in a wide array of applications, 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.

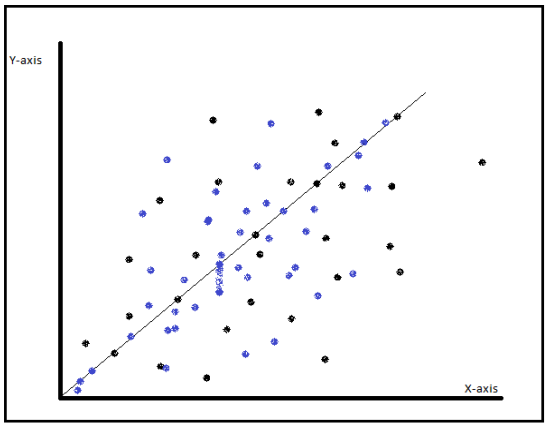

Fig.3 K-Mean clustering

II. LITERATURE REVIEW

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 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. In this approach, software architecture decomposition by clustering techniques with focus on functional requirements and attributes. 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 show that A-KNN algorithm is competitive with two Agglomerative Clustering techniques. it provide valuable information for software designer [1]. 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 neighbor 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 [13] they present study k -Nearest Neighbor classification method, have been studied for economic forecasting. Due to the effects of companies' financial distress on stakeholders, financial distress prediction models have been one of the most attractive areas in financial research. In recent years, after the global financial crisis, the number of bankrupt companies has risen. Since companies' financial distress is the first stage of bankruptcy, using financial ratios for predicting financial distress have attracted too much attention of the academics as well as economic and financial institutions. Although in recent years studies on predicting companies' financial distress in Iran have been increased, most efforts have exploited traditional statistical methods; and just a few studies have used nonparametric methods. Recent studies demonstrate this method is more capable than other methods. In this paper [4] many different software clustering algorithms have been developed with its properties, qualities and restrictions. These algorithms used for specific software systems, but the question of how to choose a clustering algorithm that is best suited for a particular software system.. In this paper, it provides a method for the selection of a software clustering algorithm for particular needs. Software clustering is a domain that has been developing for a long time. Several software clustering algorithms have been proposed. In this paper, focus only on software clustering algorithms that gather software components into subsystems that are important to someone attempting to understand the software system. The result of a software clustering algorithm is called a decomposition of the software system. 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. It also provides the method for the enhancement of existing software algorithms. This approach allows the evaluation of the new algorithm in earlier stages before its implemented.

III. KNN AND A-KNN APPROACHES

3.1 KNN Approach:

K nearest Neighbour 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 [5]. Most of the algorithm doesn't obey theoretical assumptions. It is also a lazy algorithm that does not use the training data points to do any generalization. It does not discard non support vectors like SVM. 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. Since the points are in feature space and has some distance. 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 decide 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 [6].

KNN approach is basically consisting of following steps:

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

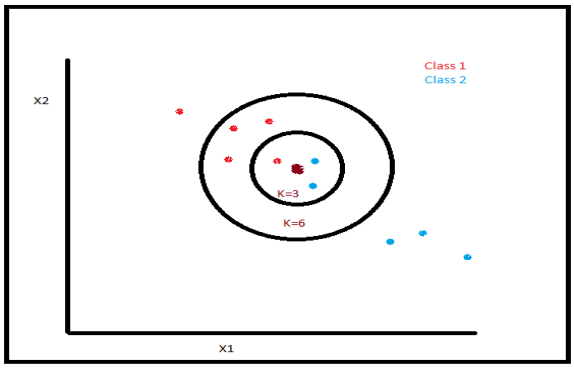

Fig. 4 KNN Approach

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. A-KNN algorithms initially consider each entity as a cluster. Each identity is labeled with a unique identifier representing the cluster identity [9]. In the

second iteration, assume that K=3, here K is the number of nearest neighbors (NN) to be selected by the user. The algorithm selects the K=3 nearest neighbors to the entity that will be clustered, and then checks labels. When two clusters out of the three clusters have the same label, the algorithm labels the current entity with the same label of those two entities. However, if the three entities have different labels, the algorithm labels the current entity with the same label of the closest entity that is nearest neighbour [10]. The algorithm repeats the clustering process until no more changes occur in the clustering tree. Then the algorithm outputs one cluster at the highest level of the hierarchy. The similarity matrix in AKNN is calculated only once. In adaptive K-mean, the number of elements in each cluster decides the number of comparison for each cluster.

IV. PROPOSED METHODOLOGY

FPA is an ISO organized which is used to find out the functional size of the system. The functional size refers the amount of functionality that is important. FPA express the size of the information in a number of function units. So its measurement unit is function units. A simplifies function point can be used to estimate the project size or length and team size. Function point estimation is performed after design creations. It required lots of judgment or empirical knowledge for accurate estimation. The functions are clustered to estimate the importance of the functions. In this paper, we have proposed a technique on A-K-NN clustering algorithm to cluster the functions on the basis of probability of the functions. The Euclidean distance has used to calculate the distance between the functions. On the basis of least distance, the functions have sent into particular cluster. In this work, we have enhanced Euclidean distance formula based on normalization to increase the efficiency of the A-KNN clustering. The enhancement has based on normalization. In the enhancement two new features have added. The first point is to calculate normal distance metrics on the basis of normalization. In second point the functions will be clustered on the basis of majority voting. The roposed technique has been implemented in MATLAB.

Algorithm

1. Iput: Dataset of Software

2. Output: Clustered Data

[row column]=size(Dataset);

1. Load dataset and define number of iterations on the dataset

2. Define 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. Plot final results of data clustering

AKNN and Enhanced AKNN algorithms are applied on the RSVP dataset to check the accuracy of both algorithms. In the last, time taken by and the accuracy of these two algorithms are compared.

V. EXPERIMENTAL RESULTS

The figure shown below demonstrates the working of simple AKNN method with the probabilistic graph. In this figure, six clusters are generated and the graph shows the probabilistic points.

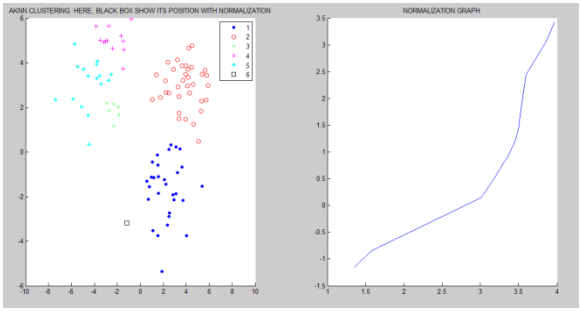

Fig 5 Simple AKNN clustering with probabilistic graph

As illustrated in the figure 1.6, the A-KNN clustering algorithm is enhanced in which normalization is applied on the RSVP dataset to improve the cluster quality. In the figure the dataset is clustered and showed on the 2D plane with different colour.

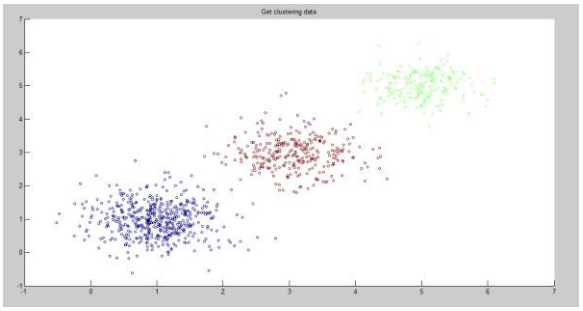

Fig 6 Function points plotted with different colours

In the figure 1.7, 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.

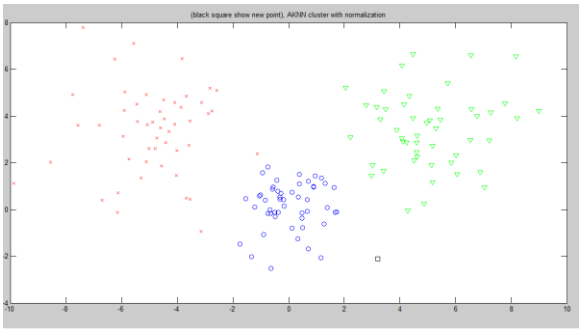

Fig 7 probabilistic point (Black Square)

In the figure 1.8, 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.

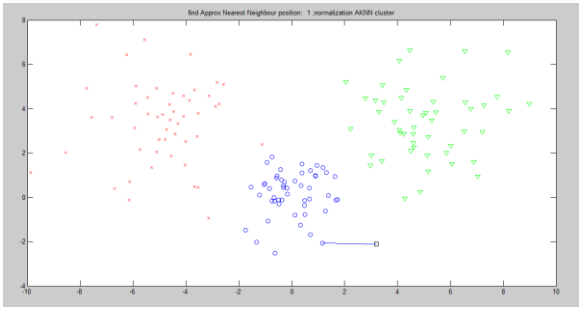

Fig 8: Plotting of data points

In the figure 1.9, 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.

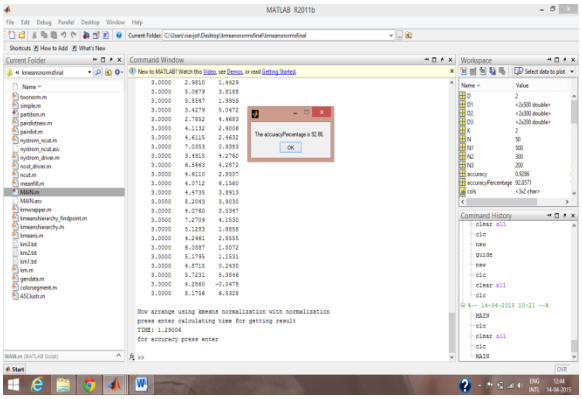

Fig 9: Plotting of data points and accuracy

5.1 Results Comparison

5.1.1 Comparison on the basis of Accuracy

The figure below demonstrates the comparison of accuracy of simple AKNN method and the enhanced normalized AKNN method. It is depicted in the figure that the enhanced AKNN clustering method provides more accuracy to create clusters as compared to simple AKNN method.

The values generated by both methods in terms of accuracy are given below in the following table:

Technique	Accuracy value
AKNN	90.00
Enhanced AKNN	92.86

Table 1 Values for comparison of Accuracy

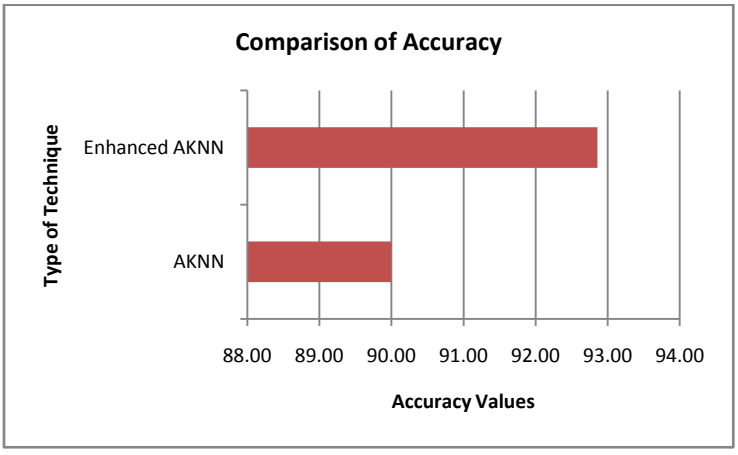

Fig 10 Graphical representation of Accuracy of AKNN and Enhanced AKNN method.

5.1.2 Comparison on the basis of total time taken

The figure below demonstrates the comparison of total time taken by simple AKNN method and the enhanced normalized AKNN method. It is depicted in the figure that the enhanced AKNN clustering method takes less time to create clusters as compared to simple AKNN method. The values generated by both methods in terms of time taken are given below in the following table:

Type of Technique	Time Taken (in secs)
AKNN	1.01045
Enhanced AKNN	1.00307

Table 2 Values for comparison of total time taken

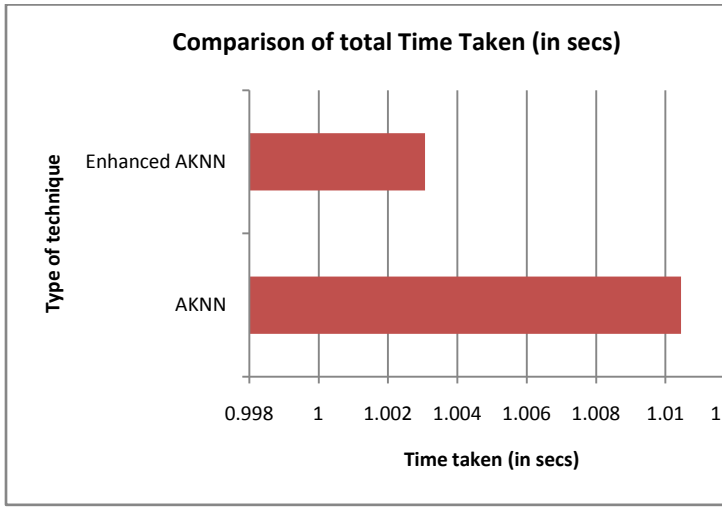

Fig 11 Graphical representation of time taken by AKNN and Enhanced AKNN method

VI. CONCLUSION

Software engineering is sequence of steps to produce the software from initial stage to its final stage. Software architecture is important part of the software development process. Software Architecture is most important factor for the development of complex and large software system. Architecture decomposition decreases the software complexity. Several clustering techniques are studied. A-KNN is producing the best result than other clustering techniques. A-KNN has used to decomposition of software Architecture with enhancing the Euclidian distance formula based on normalization that is increase the efficiency of A-KNN clustering technique. The proposed enhancement has implemented for decomposition of software architecture focus on functional requirements and attributes.

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] Pawlak. Z. Rough Sets International Journal of Computer and Information Sciences, (1982), 341-356.

[6] Pawan Lingras, Chad West. Interval set Clustering of Web users with Rough K-Means, submitted to the Journal of Intelligent Information System in 2002.

[7] Yeung K.Y, Haynor D.R, Ruzzo W.L. Validating clustering for gene expression data. Bioinformatics. 2001.

[8] Abdulaziz Alkhalida, Mohammad Alshayebb, Sabri Mahmoudb," Software refactoring at the function level using new Adaptive K-Nearest Neighbor algorithm",2010.

[9] R. Braden, L. Zhang, S. Berson, S. Herzog and S. Jamin, "Resource ReSerVation Protocol" (RSVP), 1997.

[10] Zhang Y. , Mao J. and Xiong Z.: An efficient Clustering algorithm,In Proceedings of Second International Conference on Machine Learning and Cyber netics, November 2003.

[11] Felfernig, A. Salbrechter, "Applying Function Point Analysis To Effort Estimation in configuration Development, IJEST, Vol. 4, issue 5, 2003

[12] Timothy C. Lethbridge and Robert Laganiere, "Object-Oriented Software Engineering".

[13]. Jaswinder Kaur, Satwinder Singh, Karanjeet Singh Kahlon, Pourush Bassi " Neural network-a novel technique for software effort estimation" International Journal of Computer Theory and engineering, 2010

[14] Heena Kapila, Satwinder Singh, "Analysis of CK metrics to predict software fault-proneness using bayesian inference" International Journal of Computer Applications, 2013.

[15] Satwinder Singh, KS Kahlon, "Effectiveness of encapsulation and object-oriented metrics to refactor code and identify error prone classes using bad smells", ACM SIGSOFT Software Engineering Notes, 2011