A Concept of A-KNN Clustering in Software Engineering

Poojia Gupta
Student, Computer and Science Department
Chandigarh University, Gharaun, Mohali, Punjab, India

Gurpreet kaur
Assistant Professor, Computer and Science Department
Chandigarh University, Gharaun, Mohali, Punjab, India

ABSTRACT
In Software Engineering, software decomposition plays an vital role, to increase the quality of the cluster and to handle it properly the software designer decompose the software architecture. The software decomposition is done in various ways and each way provides the different result on same dataset. In software engineering research a new technique has been investigated in which software decomposition is done by using clustering techniques. This paper presents the review of various clustering algorithms and on reviewing various algorithm it concludes that the A-KNN clustering algorithm is the efficient algorithm in terms of accuracy and complexity.

General Terms
Clustering Algorithms, Software Decomposition

Keywords
Software Architecture, Clustering, Algorithms, K-mean , KNN , A-KNN

1. INTRODUCTION
Clustering in Software Engineering: As we entered into 20th Century, the software clustering plays an important role in the field of software engineering. As we look behind the 20th century the developers have needed to prove the correctness of their software [1]. In addition to this, it is very difficult to understand the legacy systems because size and complexity of the system are constantly increasing with the time and also the maintenance of that system is too costly and quite difficult. The new approach came into existence that can help the developer to understand the large, complex software system by automatically reconstructing the system, in other words we can says that breakdown the large system into small meaningful subsystems that are easy to understand and maintain. This new approach is called software clustering [2]. The main aim of clustering is to produce a system that is highly cohesive and less coupled. 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 alike and objects in different clusters are not alike. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. 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. 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. This paper is organized as follows: Section II includes a brief overview of clustering in software engineering. Section III describes the related work. Section IV describes the Basic clustering Algorithms. Section V presents the conclusion.

2. HOW CLUSTERING IS RELATED TO SOFTWARE ENGINEERING
The main aim of clustering is to produce a software system that is highly cohesive and less coupled. Various types of clustering are discussed here:

2.1 Hierarchical Based clustering:
This type of clustering is usually applied with small dataset (less than 100 observations). In which clusters create hierarchy of clusters that are depicted in the form of tree-structures also called as dendogram [3]. This clustering technique classifies into two categories:

- Agglomerative Clustering
- Divisive Clustering

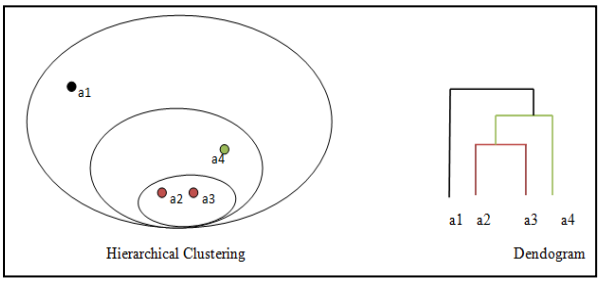

Fig 1: Concept of Hierarchical Clustering [3]

Agglomerative Clustering: It is also called as Bottom-up Clustering Approach. Start from the leave node merges two closest clusters according to distance measure at each iteration. This procedure will be repeated until there is only cluster of high quality is obtained. It is divided into three Categories:

SLINK (Single Linkage) Algorithm: also called as minimum method or Connectedness. In which compute the shortest distance between any two object of clusters, whereas one object from one cluster and other object of another cluster.

$$\min \{d(a,b) : a \epsilon X , B \epsilon Y\}.$$

CLINK (Complete Linkage) Algorithm: also called as maximum or diameter method. In which compute the farthest distance between any two object in the cluster, whereas one object from one cluster and other object of another cluster.

$$\max \{d(a,b): a \in X, b \in Y\}.$$

WPGMA (Weighted Pair Group Method using Arithmetic average): also called as average-linkage clustering. In which the distance between two clusters is taken as average of distance between all pairs of object in the two clusters.

$$\frac{1}{|X|.|Y|} \sum_{a \in X} \sum_{b \in Y} d(a,b)$$

Divisive Clustering: It is also called as Top-down clustering approach. Start from the root node and subsequently decomposing the group into smaller ones until every object falls in one cluster. This approach follows the principle of divide-and-conquer algorithm.

2.2 Graph-Based Clustering:

Basically graph-based clustering are of two type:

- Connected clusters
- Clique.

In this type , the clusters are depicted in the form of graphs where as nodes are objects and edge represent connection among objects, then that type of clusters called as connected clusters; i.e a group of objects are connected with in the cluster not outside the clusters. Continuity-based clusters are the example of graph-based clusters. In which the clusters are mostly irregular in shape and problem occur when clusters contain some extent of noise. Clique is defined as set of nodes in a graph which are interconnected to each other. The graph-based clusters need to be globular.

2.3 Density-based clustering

In this type of clustering, the neighboring objects are clustered on the basis of density condition. The grouped objects are surrounded by a low-density region. These types of clusters are often employed when clusters are present in irregular shape and some amount of noise and outlier is present in the clusters.

2.4 Partitional-based Clustering

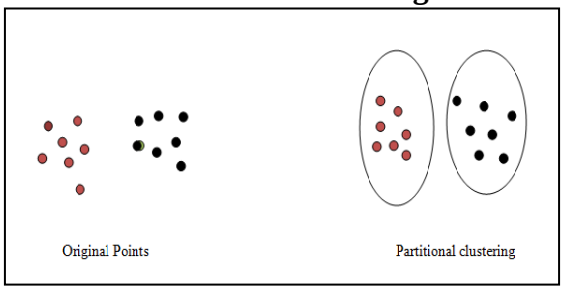

Fig 2: Concept of Partitional Clustering [3]

It creates set of K partitions, then use the relocation technique to improve the partitions by moving object from one group to another group. The partitional based clustering is the most important clustering in software engineering. In partitional based the most important type of clustering is K-mean Clustering. In k-mean clustering the basic idea is to categorize the data into k clusters whereas k is the input parameter enumerated in advance through iterative relocation method that converges to the innate minimum. It consists of two distinct periods: Early period is to ascertain k centers at random one for each cluster. The subsequent period is to ascertain distance amid data points in Dataset and the cluster centers and allocating the data point to its nearest cluster. Euclidean distance is usually believed to determine the distance. When all the data points are encompassed in little clusters an early grouping is done. New centers are next computed by seizing the average of points in the clusters. This is done because of inclusion of new points could lead to a change in cluster centers. This procedure of center updation is iterated till a situation whereas centers do not notify anymore or criterion the function becomes minimum. The basic concept of partitional based clustering is described in figure 2. Where first consider original points and after that partitional based clustering applied on that data points.

Table 1: Basic Clustering Techniques and their examples

Types of Clustering	Examples
1. Partitional Clustering	K-mean PAM (Partitioning Around Mediods)
2. Hierarchical Clustering	BIRCH CURE ROCK
3. Density-based Clustering	DBSCAN DENCLUE
4. Grid-based Clustering	STING Wave cluster

3. RELATED WORK

Estimating of Software Quality with Clustering Techniques [3]: in this paper different models of clustering technique are compared: K-mean, Hierarchical Agglomerative clustering and Fuzzy C-mean clustering. In this paper basically two types of clustering are used: Distance-based clustering and conceptual clustering. In Distance-based clustering the objects are clustered on the basis of distance between the cluster, those clusters have minimum distances are clustered. On the other hand in conceptual based clustering the two or more objects are clustered if they share common concept to all the objects. When these techniques are applied on the two NASA software project and concluded that the effective result is produced by Fuzzy C- mean clustering because it gives the cluster of good quality as compared to k-Mean clustering technique.

A Heuristic Approach to Solving the Software Clustering Problem [4]: Introduced a metaheuristic search based technique to solve a complex and large software clustering problem. In which algorithm and tool to cluster large software system in a well efficient manner was created. The automatic clustering algorithm called "Bunch" is used. Bunch used the fully-automatic, semi-automatic and manual clustering features. The most software clustering algorithm produced same result on the given input but Bunch clustering algorithm uses randomized heuristic search technique in which on giving the same input on same run it gives the different result. A tool called "CRAFT" that uses the Bunch's software

clustering algorithm. The output of this tool is to generate reference decomposition and various similarity measures like EdgeSim and Meci were developed to calculate and compare various software clustering algorithm. The conclusion is that it gives the optimal solution to solve the NP- hard problem.

ABACUS: Mining Arbitrary Shaped Clusters from Large Datasets based on Backbone Identification [5]: Proposed a robust and shape-based clustering algorithm. This is mainly identifying the backbone of the clusters; the backbone includes the set of core points within the cluster. ABACUS operates in two steps: in the first step , it identify the backbone of each cluster by means of iterative process like globbing (or merging of points) and movement of point operation and in the 2nd stage , Once the backbone is identified, it is very easy to recover the final set of clusters from it. The globbing and movement point operation have two main benefits: 1) it remove the noise points from the data set 2) Dataset size is concise. The conclusion is that the ABACUS clustering algorithm gives the cluster of good quality as compared to other arbitrary shaped clusters. This type of algorithm is mainly used in image processing.

Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data [6]: Introduced a novel clustering technique called a Shared Nearest Neighbor (SNN).The SNN clustering algorithm find clusters of various shapes, size and densities even when the noise and outlier is present. This algorithm also handles the large amount of data and find out the no. of clusters from them. It can find the "dense" clusters like other clustering algorithms but it can also find the clusters of low and medium density which are of interest because in which the uniform region is "surrounded" by the high density area or non-uniform clusters. In which an example is presented in context of NASA Earth Science time series data, where SNN clustering algorithm find the clusters of different densities that was important for researcher to describe how the ocean influence the land. In contrast with DBSCAN it can find only the weaker clusters. K- mean also find the weaker cluster but from the large cluster it can find only one and cluster obtained has low quality. So at last the clusters produced by SNN clustering algorithm are of good quality as compared to others clustering algorithms.

Abdulaziz Alkhalid [7], introduced 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 neighbor 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.

Towards effective and efficient mining if arbitrary shaped Clusters [8]: proposed a CLASP (Clustering aLgorithm for Arbitrary ShaPed Clusters) an effective and efficient algorithm for mining arbitrary shaped clusters. This algorithm is perform in 3 stages: first , it automatically shrink the size of data set while effectively preserving the computational cost and shape information of clusters. Second , it adjust the positions of the representative data examples. And finally , it perform the agglomerative clustering to identify the cluster structure with the help of a mutual k-nearest neighbors –based similarity metric (novel similarity metric) called $P_{k.}$ then Extensive experiment is performed on both synthetic and real data sets , and result verifies the effectiveness and efficiency of this approach.

Software Architecture decomposition using clustering technique [10]: Introduced a Fuzzy C-mean(FCM) clustering along with three hierarchical Agglomerative clustering algorithm and adaptive k-nearest neighbor(A-KNN) algorithm. There are two issues which are generally arises in the software system decomposition when clustering technique is applying: First, it is difficult to determine the no. of clusters. Second problem is, the clustering tree doesn't show clearly how close a component is related to various clusters. In this paper, FCM technique help the software designer in determining the no. of clusters. When applying there algorithms on NASA dataset. The result showed that the A-KNN algorithm is competitive with the three agglomerative clustering algorithms. FCM provides valuable and objective information and also effective in handling the two common challenges for clustering technique.

4. BASIC SOFTWARE CLUSTERING ALGORITHMS

K-Mean Clustering Algorithm

The traditional technique of clustering is k-mean clustering. It is a type of non-hierarchical clustering. The word k-Mean was first used by MacQueen in 1967, it is one of the simplest unsupervised learning algorithm that solve the well-known clustering problem. Basic properties of K-Mean clustering are:

- There is no objective function here so it is popularly called subjective segmentation.
- There is no dependent variable, only have input parameters. i.e grouping based on independent variables.
- The segmentation develops on its own based on value of the input variables.
- That's why it is unsupervised learning because we are not trying to set the pattern it is just extracting the pattern on its own.
- It is different from decision tree because decision tree has objective function. E.g. 0/1 for the dependent variable.

Algorithm step for K-Mean clustering:

1) Firstly place k point into the space represented by the elements that are being clustered. These points represent the group of centriods.

2) Assign each element to the group that has the closest centriod.
3) Once all the elements have been assigned, recalculate the position of k centroids.
4) Repeat step 2 and 3 until no more change occur.

The main aim of K-Mean clustering algorithm is to minimize the objective function; here the objective function is squared error function [3].

KNN (Nearest Neighbor) Clustering Algorithm

KNN is a classification algorithm was first proposed by T.M. Cover P.E. Hart. It is one of the top ten algorithm in data mining which has been applied in various field like pattern recognition, text classification, etc. In which space is partitioned into regions by the locations and labels of the training samples [7].

Algorithm step for KNN clustering:

1) Determine the value of k.
2) Then calculate the distance between the new point and all the training data.
3) Find out the distance and determine the k nearest neighbor based on the k^{th} distance.
4) Gather all the categories of those neighbors and find out the category based on the highest confidence [9].

A-KNN Clustering Algorithm

A-KNN depicts the Adaptive K- Nearest Neighbor clustering. It is the extension of K-NN clustering. The main advantage of A-KNN clustering over other clustering technique is it reduce the amount of computation because in which the similarity matrix is calculated only once [10].

Algorithm steps for A-KNN clustering:

1) Clusters the data into k groups where k is predefined.
2) Select k points using normalization technique.
3) Assign objects to their closest cluster according to the Euclidean distance function.
4) Calculate the centriod or mean of all objects in each cluster.
5) Repeat setps 2,3 and 4 until the same points are assigned to each cluster in consecutive rounds.

5. CONCLUSION

In this paper we have reviewed all the clustering algorithm and conclude that the A-KNN clustering algorithm is better one because it has less complexity and give more accurate clusters. Also we discussed that in order to improve the quality of cluster, we can use of various clustering techniques like hierarchical clustering, divisive clustering, etc. In software engineering software clustering is used to simplify software maintenance and improve program understanding. The A-KNN clustering technique has less computation cost because it calculates the similarity matrix only once. So it requires less execution time, by speed up the algorithm as compared to traditional ones.

6. REFERENCES

[1] Bozhikova, Violeta. "Using Clustering to Achieve Quality Software Structure."Electronics Technology, 2006. ISSE'06. 29th International Spring Seminar *on*. IEEE, 2006

[2] Andritsos, Periklis, and Vassilios Tzerpos. "Information-theoreticsoftware clustering." Software Engineering, IEEE Transactions on 31.2 (2005): 150-165.

[3] Gupta, Deepak, Vinay Kr Goyal, and Harish Mittal. "Estimating of Software Quality with Clustering Techniques." Advanced Computing and Communication Technologies (ACCT), 2013 Third International Conference on. IEEE, 2013.

[4] Mitchell, Brian S. "A heuristic approach to solving the software clustering problem." Software Maintenance, 2003. ICSM 2003. Proceedings. International Conference on. IEEE, 2003.

[5] Chaoji, Vineet, et al. "ABACUS: Mining Arbitrary Shaped Clusters from Large Datasets based on Backbone Identification." SDM. 2011.

[6] Ertöz, Levent, Michael Steinbach, and Vipin Kumar. "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data." SDM. 2003.

[7] Alkhalid, Abdulaziz, et al. "Software Architecture Decomposition Using Clustering Techniques." Computer Software and Applications Conference *(COMPSAC)*, 2013 IEEE 37th Annual. IEEE, 2013

[8] Huang, Hao, et al. "Towards effective and efficient mining of arbitrary shaped clusters." Data Engineering (ICDE), 2014 IEEE 30th International Conference on. IEEE, 2014.

[9] Shtern, Mark, and Vassilios Tzerpos. "Clustering methodologies for software engineering." Advances in Software Engineering 2012 (2012).

[10] Alkhalid, Abdulaziz, Chung-Horng Lung, and Samuel Ajila. "Software architecture decomposition using adaptive K-nearest neighbor algorithm."Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on. IEEE, 2013.

[11] Bauer, Markus, and Mircea Trifu. "Architecture-aware adaptive clustering of OO systems." Software Maintenance and Reengineering, 2004. CSMR 2004. Proceedings. Eighth European Conference on. IEEE, 2004.

[12] Doval, Diego, Spiros Mancoridis, and Brian S. Mitchell. "Automatic clustering of software systems using a genetic algorithm." Software Technology and Engineering Practice, 1999. STEP'99. Proceedings. IEEE, 1999.

[13] Desai, Narayan, et al. "Component-based cluster systems software architecture a case study", Cluster Computing, 2004 IEEE International Conference on. IEEE, 2004.