International Journal of Computers & Technology
Volume 5, No. 2, May -June, 2013, ISSN 2277-3061

Maintainability Evaluation of Object-Oriented Software System Using Clustering Techniques

Astha Mehra[1], Sanjay Kumar Dubey[2]
1, 2 Amity University, Sec-125, NOIDA (U.P.) India
1astha90@gmail.com, 2skdubey1@amity.edu

ABSTRACT

In today's world data is produced every day at a phenomenal rate and we are required to store this ever growing data on almost daily basis. Even though our ability to store this huge data has grown but the problem lies when users expect sophisticated information from this data. This can be achieved by uncovering the hidden information from the raw data, which is the purpose of data mining. Data mining or knowledge discovery is the computer-assisted process of digging through and analyzing enormous set of data and then extracting the meaning out of it. The raw and unlabeled data present in large databases can be classified initially in an unsupervised manner by making use of cluster analysis. Clustering analysis is the process of finding the groups of objects such that the objects in a group will be similar to one another and dissimilar from the objects in other groups. These groups are known as clusters. In other words, clustering is the process of organizing the data objects in groups whose members have some similarity among them. Some of the applications of clustering are in marketing -finding group of customers with similar behavior, biology- classification of plants and animals given their features, data analysis, and earthquake study -observe earthquake epicenter to identify dangerous zones, WWW -document classification, etc. The results or outcome and efficiency of clustering process is generally identified though various clustering algorithms. The aim of this research paper is to compare two important clustering algorithms namely centroid based K-means and X-means. The performance of the algorithms is evaluated in different program execution on the same input dataset. The performance of these algorithms is analyzed and compared on the basis of quality of clustering outputs, number of iterations and cut-off factors.

General Terms

Software Engineering, Data Mining

Indexing terms

Maintainability, Clustering, Software Metrics, Quality Models, K-Means, X-Means

Academic Discipline And Sub-Disciplines

Engineering

SUBJECT CLASSIFICATION

E.g., Mathematics Subject Classification; Classification

TYPE (METHOD/APPROACH)

Experimental

1. INTRODUCTION

For any software product, maintainability is considered to be one of the most important phase of the software development life cycle. Maintainability is treated as a quality attribute because the cost of the work required to perform the maintenance activities on any software product constitutes for the largest cost in present scenario of software development. The ISO/IEC 9126 [1] standards defines maintainability as the capacity of the software product to be modified, including corrections, improvements or adaptations of software to change in environment and in requirements and functional specifications. According to [2]-[3], maintenance is a necessary task but at the same time it is a complex section of the entire software life cycle usually constituting of 50-70 proportion of total effort allocated to a software system. Quality of the source code and quality of internal attributes (such as coupling, inheritance, etc) largely impacts the maintainability of the software product. It is very necessary to meet the customer's requirements and they can be met through good maintainable software. Software maintainability can be basically analyzed through three ways namely qualitative analysis, experiences of experts and quantitative measurement. Software maintainability helps us to improve the existing features of the software product, to identify and correct faults and errors, and provide portability to the software to move to different working environment.

The advantage of the data mining technology is that it can deal with large amount data of data efficiently and can extract the useful information out of it. This attribute of data mining has been helpful in improving the software maintenance [4]-[6]. If we are able to analyze the maintainability of the software product as early as at the design level then it will enable the designers and the maintainers to make necessary changes in the architecture of the software system which would provide better performance. This will further lead to lowering the maintainability cost [7].

Clustering is defined as the process with the help of which we can group the data objects which are similar to each other as one cluster. Dissimilarity is observed in the data objects belonging to different clusters. Apart from the use of clustering

International Journal of Computers & Technology
Volume 5, No. 2, May -June, 2013, ISSN 2277-3061

technique in general applications such as image processing, data analysis, marketing, WWW, etc; it is also used to evaluate and improve the maintainability of the software system. For this purpose we have various clustering algorithms. But in this paper we have compared and analyzed the effect of K-means clustering and X-means clustering.

The rest of the paper has been organized as follows. In section 3 we have discussed about clustering in details. Section 4 describes details about the K-means clustering algorithm and section 5 deals with X-means clustering algorithm. In section 6 we have experimental results and discussions on K-means and X-means. Finally section 7 concludes our paper.

2. CLUSTERING

Clustering analysis is the process of finding the groups of objects such that the objects in a group will be similar to one another and dissimilar from the objects in other groups. These groups are known as clusters. We can also say that the inter cluster distance should be maximized and the intra cluster distance should be minimized. In other words, data points of a dataset in same cluster are highly related to each other whereas the data points in different clusters are not related to each other. Clustering is sometimes also known as data segmentation as it is used to partition huge data in segments called as clusters. Clustering is applicable for issues such as data analysis where very little information is available and the certain assumptions need to be made, image processing, unsupervised learning, pattern recognition, biology, etc. But, there are some typical requirements of clustering in data mining such as ability to deal with noisy data, scalability, ability to deal with different type of attributes, discovery of clusters with arbitrary shapes, etc. With the help of clustering approach we can identify co-relation among the attributes of the dataset and the distribution pattern of the particular dataset.

The advantages of clustering based approach to improve maintainability is that these processes are adaptable to change and they also tell us about the features which helped in distinguishing the groups or clusters. Clustering can be performed on the dataset in the following ways [8]:

- Partitioning Methods
- Hierarchical Methods
- Density Based Methods
- Grid Based Methods
- Model Based Methods
- Clustering High Dimensional Data
- Constrain Based Clustering

1. Partitioning Methods: In this method we divide the dataset of n objects into k partitions or groups and each partition represents a cluster and k <= n. We have to make sure that each group must contain minimum one object and each object must belong to exactly one group. If we have to make k clusters or groups then partitioning method will create an initial partitioning. It then makes use of an iterative relocation technique and relocates objects from one group to another so as to improve the partitioning. Examples are the k-means algorithm and the k-medoids algorithm.

2. Hierarchical methods: It creates a hierarchical decomposition of the given set of data objects. Based on the formation of hierarchical decomposition, a hierarchical method can be divided into two types being agglomerative or divisive. The agglomerative approach, also called the bottom-up approach, starts with the process of each object forming a separate group. It then successively merges the objects or groups depending upon their closeness to one another. This process continues until all of the groups are merged into one that is until the topmost level of the hierarchy is formed, or until a termination condition holds. The divisive approach which is also known as the top-down approach, starts when all the objects in the same cluster. After every iteration or step, a cluster is split up into smaller clusters. This process continues until each object is in one cluster, or until a termination condition holds.

3. Density-based methods**:** In this method we continue expanding the given cluster till the time density in the neighborhood exceeds some threshold. It is used to filter out noise (outliers) and discover clusters of arbitrary shape. Examples are DBSCAN and its extension, OPTICS etc.

4. Grid-based methods**:** In this method we quantize the object space into a finite number of cells that form a grid structure. The clustering operations are performed on the grid structure which we have formed. The advantage of this approach is that it has fast processing time because it dependents only on the number of cells in each dimension in the quantized space and not on the number of data objects. STING is a typical example of a grid-based method.

5. Model-based methods: In this approach we construct a hypothetical model for each of the clusters. After constructing the model we find which model will best fit the data in the dataset. It may locate clusters by constructing a density function that reflects the spatial distribution of the data points.

6. Clustering high-dimensional data**:** This is considered as one of the important task in clustering analysis because many applications require us to analyze objects which contain a large number of features or dimensions. As the number of dimensions increases, the data become increasingly sparse and the distance measurement between pairs of points become meaningless. Also the average density of points anywhere in the data is likely to be low. Therefore, a different clustering methodology needs to be developed for high-dimensional data. CLIQUE and PROCLUS are two influential subspace clustering methods.

7. Constraint-based clustering: In this type of clustering we incorporate user-specified or application-oriented constraints. A constraint helps to tell us that what the user's expectations are or describes properties of the desired clustering results. It also provides an effective means for communicating with the clustering process.

International Journal of Computers & Technology
Volume 5, No. 2, May -June, 2013, ISSN 2277-3061

Most commonly we use the first three approaches that is partition method, hierarchal method and the density based methods.

3. K-MEANS

K-means clustering is a simplest partition method based algorithm to classify or group objects based on their attributes and features into K number of groups. It is centroid based partition technique where K stands for the number of clusters or group desired by the user. It is basically unsupervised learning algorithms that solve the well known clustering problem [9]. The aim in the K-means clustering is that we try that the objects or the data points in the same clusters should be as close to each other as possible and the objects or the data points in different clusters should be as separated from each other as possible. In k-means clustering each cluster is identified by its centre point that is the centroid.

K-means makes use of loops to classify or group the data items into the specified number of K groups. K-means is considered as a non-hierarchical clustering technique. It is based on the concept of iterations. In K-means clustering we begin the iterative process by first finding out the initial centre point i.e. the centroid of each group. The selection of this centre point is done randomly by selecting one data point from each of the group. Then we assign each data point to that group which is closest to it by calculating the Euclidean distance between each data point to each centroid. Each data point is allocated to the nearest group. After this each group will find its new centre point by calculating the centroid to replace the initial one and the process of computing the Euclidean distance and then assigning the data objects to the nearest group continues. These iterations stops when the data objects do not change their group i.e. each group attains stability or the requested number iterations have been performed.

In K-means we have a function which tells us the aggregate dissimilarities of the clusters or groups. This function is called square error function and is defined by [10]:-

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

J= sum of square error for all objects in data set.

$x_i^{(j)}$= Point in space representing a given object.

c_j= mean of cluster ci.

3.1 The K-means Algorithm:

1. Specify number of groups and select initial centroid of each group.
2. Calculate Euclidean distance for each data object and centroid. Assign the data objects to the nearest centroid.
3. Calculate distance's mean of every data member and own centroid to define new centroid in each group.
4. Repeat steps 2 and 3 until each the data objects do not change their group or specified number of iterations have been performed.

4. X-MEANS

Even though K-means has certain advantages but there are three major drawbacks of K-means algorithm. First, it is slow and takes large amount of time to complete each iteration. Second, the number of clusters or groups has to be pre-defined by the user. Third, whenever we execute this algorithm, because of fixed value of k, it always finds worse local optima as compared to the situation when it can dynamically change k. These problems can be resolved to some extend by the help of X-means algorithm which helps in efficient estimation of the number of clusters. The building block of X-means algorithm is the concept of blacklisting. In blacklisting [13], for a given region, we maintain a list of specific centroids that needs to be considered. Blacklisting is a very fast process and it can handle large number of centroids.

In X-means algorithm the output tells us about the number of classes and their parameters. In this algorithm the entire data is split on some parameters (described below) and K-means is applied on each of them. After the implementation of K-means, X-means algorithm makes decision that in order to better fit the data, the subsets of the centroid should be split or not. The decision whether to split or not is done by computing the Bayesian Information Criterion (BIC). Bayesian Information Criterion, in statistics, helps us to select a model from a set of models. It is based on likelihood function. Applications of BIC include models which are based on maximum likelihood and model identification in time series. With respect to BIC, X-means has been considered to produce better clustering on real-life data as well as synthetic data.

The splitting of data can be basically done in two ways [13]. First approach is "one at a time". In this technique we can select any one centroid, produce a new centroid and implement k-means to completion. After this we check if the resultant model scores better or not, if the resultant model scores better then we accept it otherwise we return to the previous structure. In this way all the centroids will be tested and the best result will be produced. But this will be an extremely time-consuming and expensive operation. Also, it will need $O(K_{max})$ steps till the time X-means is completed. Second approach is "try half the centroids". In this technique, with the help of some heuristics criterion we choose half the centroids to be split. After splitting, we execute k-means algorithm on them and check if the resultant model is better than the original or not. If it is better then we accept the model else we return to the previous structure. In this technique, we

International Journal of Computers & Technology
Volume 5, No. 2, May -June, 2013, ISSN 2277-3061

need only O(log K_{max}) steps till the time X-means is completed. But the problems lies in distortion produced due to centroid, which heuristics criterion should be chosen and size of the region owned by the centroid.

4.1 The X-means Algorithm

It consists of two basic operations which are repeated until completion. The operations are as follows:-

1. Improve-Params:- this is a very simple operation. It just consists of execution of K-means to convergence.
2. Improve-structure:- this operation tells us whether a new centroid should be formed or not. It basically involves splitting the centroid into two, if required.
3. If k > Kmax, the we stop and submit the best scoring model which was found during the search. Else, goto 1.

5. EXPERIMENTAL DISCUSSIONS

We have made use of WEKA tool to perform the experiment. Weka has tools which help in classification, data pre-processing, regression, clustering, association rules and visualization. The experimental test is carried on QUES class's data [14]. It consists of 71 classes and 11 attributes. These attributes helps us to measure the performance of the classes. The 11 attributes describing these 71 classes are:-

1. Depth of Inheritance Tree (DIT) - For each class it provides measure of level of inheritance from objects hierarchy top.
2. Number of Children (NOC) - It provides the measure of the number of immediate children of the class.
3. Message Passing Coupling (MPC) - The complexity of message passing amongst classes is measured by it.
4. Number Of Methods (NOM) – It provides the measure of number of local method which are there in a class.
5. Response of Class (RFC) – It counts set of all the methods that can be invoked as a response to all the methods which are accessible within the class hierarchy.
6. Data Abstraction Coupling (DAC) – The coupling complexity caused by ADT's is measured by this attribute.
7. Lack of Cohesion Method (LOCM) – It checks if all the methods of a class work properly together and in order so as to achieve single and well defined purpose.
8. Weight Method per Class (WMC) – It provides the sum of the complexities of the methods in a particular class i.e. the overall complexity.
9. SIZE 1 - It provides us with the measure of total number of semi-colon in a class.
10. SIZE 2 – It provides us with the measure of number of attributes and number of methods in a class.
11. Change – The number of lines changed per class gives us the measure of maintenance effect.

In WEKA, when we open the QUES class's data then WEKA explorer shows all the details of the data set in the format as shown in fig(1)

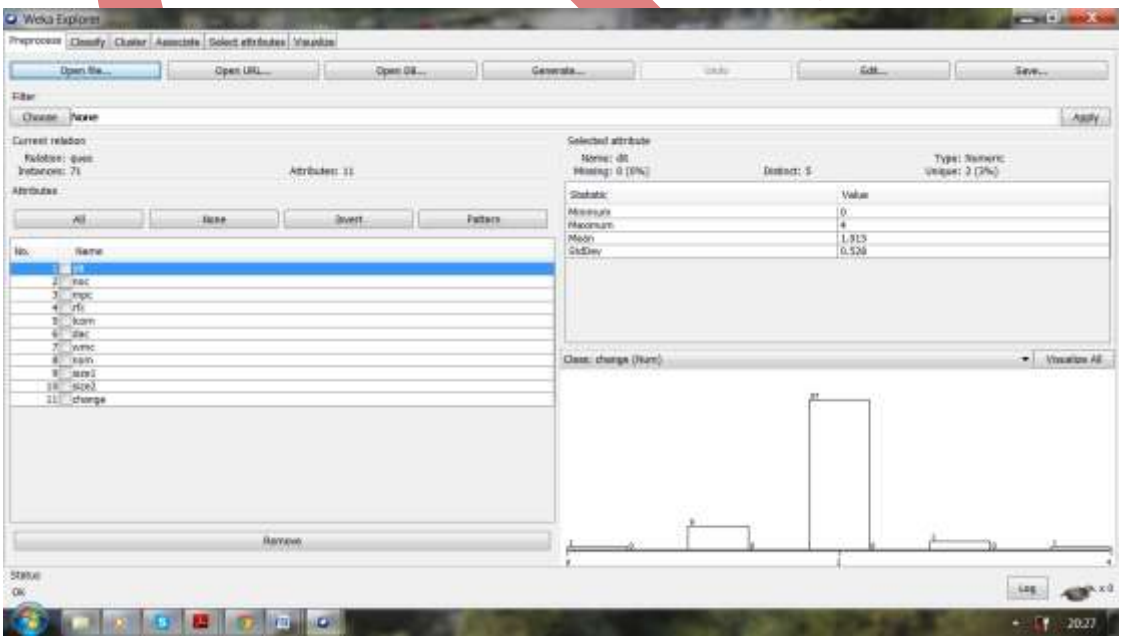

Figure 1: Attributes of QUES data set.

It shows that the relation name i.e. the name of the data set is QUES. It contains a total of 71 instances and 11 attributes. We can also remove some of the attributes for analyzing the clustering algorithm by clicking on the names of those attributes. All the 11 attributes can be visualized in a graphical form as shown in figure 2.

International Journal of Computers & Technology
Volume 5, No. 2, May -June, 2013, ISSN 2277-3061

Figure 2: Individual graphical description of attributes of QUES data set.

5.1 Implementation of K-means on QUES data set.

K-means is implemented on WEKA tool by directly choosing the Simple K-means clustering technique from the drop down menu. The inputs such as seed value, number of clusters and number of iterations is provided by the user. In our experiment we have given the following values:-

Maximum number of iteration= 10, Number of clusters= 4, Seed= 10

After the execution of K-means algorithm on this data set with the above mentioned input we get the output as shown in figure 3. In this the number of iteration performed are 9 and shows the population of instances in different clusters.

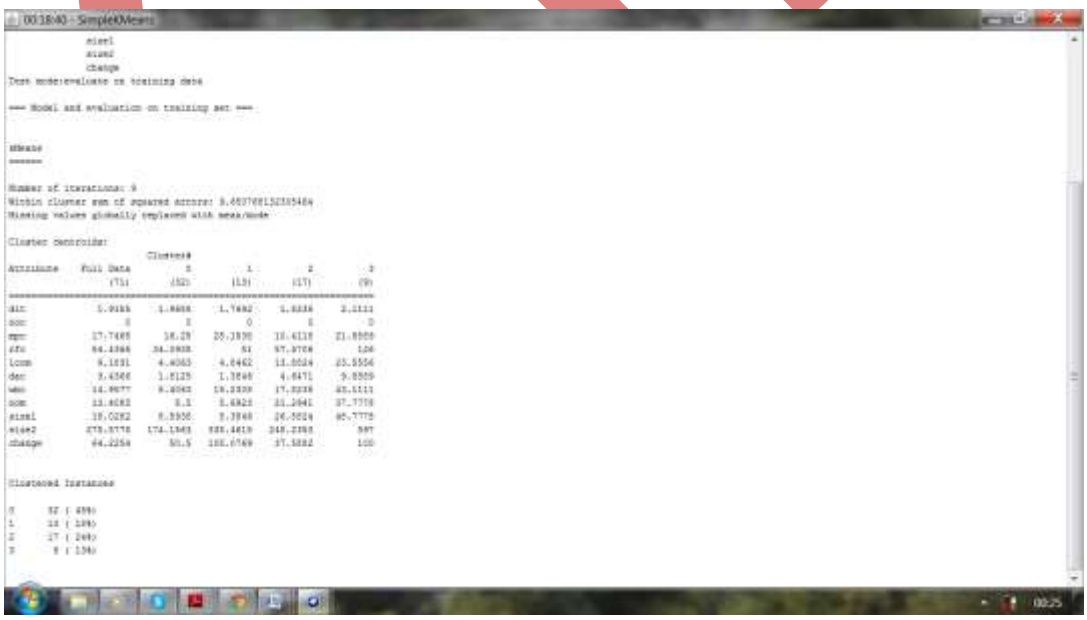

Figure 3: Detailed description of clusters in K-means.

The visualization of clusters in K-means algorithm on QUES data set can be shown by the figure 4. In the figure 4 we can see that the X-axis represents the attribute change which gives the measure of maintainability and Y-axis represents the cluster number.

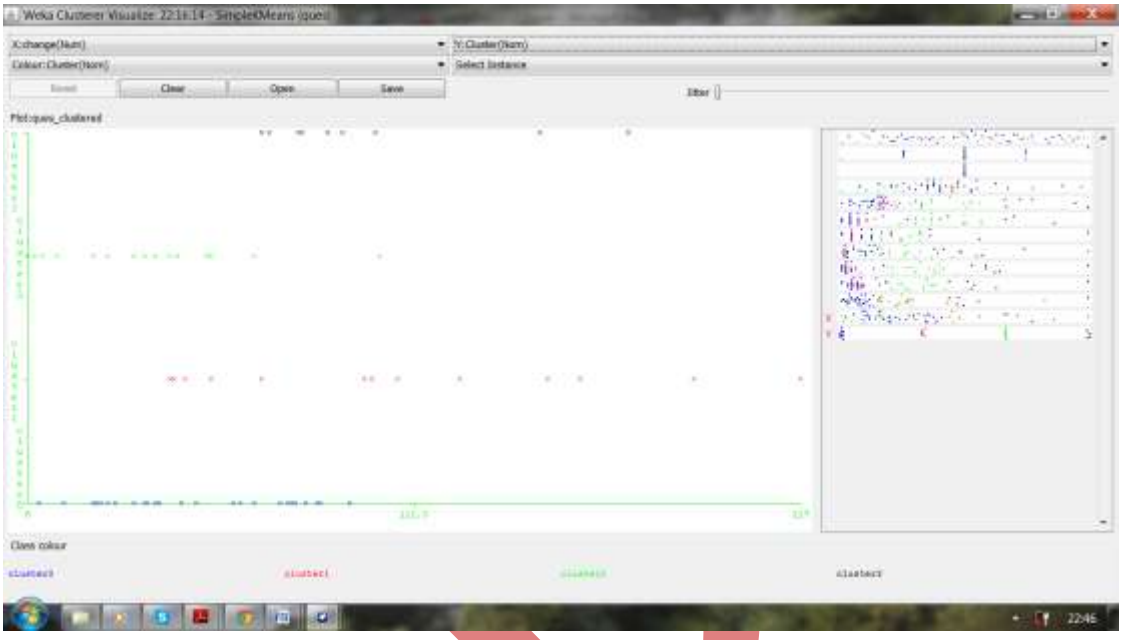

Figure 4: Graph depicting clusters and the maintainability measure of K-means.

5.2 Implementation of X-means on QUES data set

X-means is implemented on WEKA tool by directly choosing the X-means clustering technique from the drop down menu. The inputs such as seed value, minimum number of clusters, maximum number clusters, maximum k means, cut-off factor and maximum number of iterations is provided as in input by the user. In our experiment we have given the following values:-

Cut Off Factor= 0.5

Max Iterations=10

Max k-means= 1000

Max num clusters=5

Min num clusters= 4

Seed= 10

After the execution of X-means algorithm on this data set with the above mentioned input we get the output as shown in figure 5. In this the number of iteration performed are 9 and shows the population of instances in different clusters along with the BIC values, distortion and information about the splits.

Figure 5: Detailed description of clusters in X-means.

International Journal of Computers & Technology
Volume 5, No. 2, May -June, 2013, ISSN 2277-3061

The visualization of clusters in X-means algorithm on QUES data set can be shown by the figure 6. In the figure 6 we can see that the X-axis represents the attribute change which gives the measure of maintainability and Y-axis represents the cluster number.

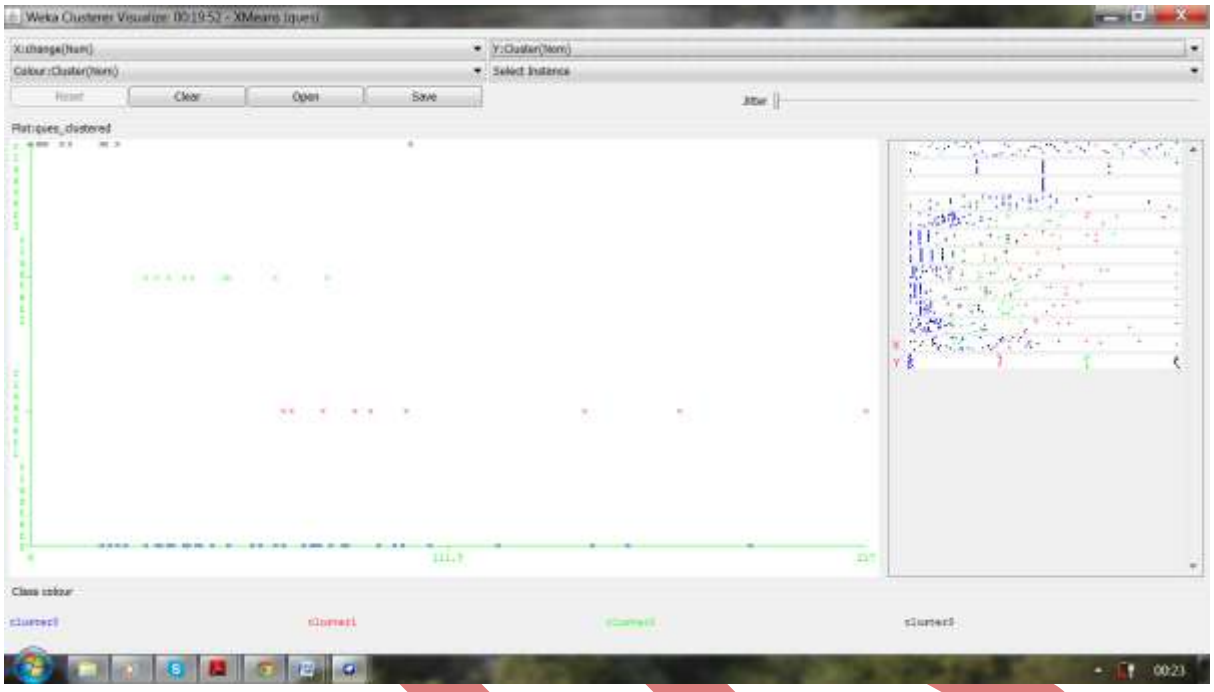

Figure 6: Graph depicting clusters and the maintainability measure of X-means.

5.3 Comparison of K-means and X-means

We can compare K-means and X-means on a number of parameters on the basis of our experiment. Firstly, we see that in K-means number of iteration performed were 9 whereas in X-means even though the requested number of iteration were same as that in case of K-means but it just performed 1 iteration to produce the result. So, X-means is faster than K-means. Secondly, we get the distortion and BIC values in case of X-means which are not produced by K-means algorithm. Thirdly, for each iteration, X-means prepared splits so that we can have better estimation of number of clusters. Then, in K-means the missing values are replaced with the mean values but in case of X-means there is no missing value. X-means also help in model selection i.e. when it decides whether the split should be performed or not. Also, there is difference between the number of instances which are present in each cluster which is represented in table 1. So, on the above made comparison, we can say that X-means algorithm is better than the conventional K-means algorithm.

Table 1. Cluster Comparison of K-means & X-means

Cluster Number	K-MEANS	X-MEANS
CLUSTER 1	32 (45%)	40 (56%)
CLUSTER 2	13 (28%)	9 (13%)
CLUSTER 3	17 (24%)	10 (14%)
CLUSTER 4	9 (13%)	12 (17%)

6. CONCLUSION AND FUTURE SCOPE

In this study we have compared the two clustering algorithm namely, K-means and X-means on various parameters for analyzing the maintainability of QUES data set. Maintainability should be analyzed at the design phase of software development life cycle itself. K-means has certain disadvantages such as the number of clusters have to pre-defined, its speed, etc but these disadvantages can be eliminated to certain extend with the help of X-means clustering technique. Experimental results also show that the X-means algorithm is better than the K-means clustering. Similar studies can be done to compare K-means and X-means on other parameters such as time elapsed, noise, etc. Various such data mining clustering techniques can be analyzed on large data sets. Moreover, these two algorithms can also be compared on

International Journal of Computers & Technology
Volume 5, No. 2, May -June, 2013, ISSN 2277-3061

some other tool such as MATLAB. Also, X-means algorithm can be compared with other data mining techniques to achieve better maintainability on the basis of number of iteration, noise, complexity, etc.

REFERENCES

[1] ISO/IEC 9126, "International technology- software product evaluation- Quality characteristics and guidelines for their use", International Standard Organization, Geneva 1991.

[2] Pigoski T.M., Practical Software Maintenance: Best Practices for Managing your Software Investment, Wiley Computer Publishing, 1996.

[3] Sommerville, Software Engineering, 6th ed., Harlow, Addison-Wesley, 2001.

[4] Anponellis, P., Antorious, D. and others, "A Data Mining Methodology for Evaluating Maintainability According to ISO/IEC-9126 Software Engineering-Product", Internet.

[5] Kanellopoulos, Y. and Others, "K-attractors: A clustering Algorithm for Software Measurement Data Analysis," 19th IEEE International Conference on Tools with A.I.

[6] Malviya, A.K. and Dutta, M., "Measuring the Maintainability of Object Oriented Systems", International Journal of Information & Computing Science, Vol. 7 and No. 2, pp. 1-12.

[7] Muthana, S., Kontogiannis, k., Ponnambalam, K. and Stacey, B. "A Maintainable Model for Industrial Software Systems Using Design Level Metrics", IEEE Software, 2000.

[8] file:///C:/Users/astha/Desktop/Project%208th%20Sem/Chapter%206%20%20%20Cluster%20Analysis%20%20%20Data%20Mining%20and%20Warehousing.htm, Internet.

[9] Kanungo Tapas, Mount David M., 2002, —A Local Search Approximation algorithm for K-mean Clustering‖, Communication of ACM.

[10] Malviya, A.K. and Dutta, M., "Maintenance Activities in Object Oriented Software Systems Using K-means Clustering Technique: A Review", Software Engineering (CONSEG), 2012 CSI Sixth International Conference

[11] Lawrence K.D., Kudyba S. and Klimberg R.K., Data mining methods and applications. USA: Auerbach Publications, 2008, pp. 83-104.

[12] Rui Xu, Donald Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks, Vol.16, No.3, May 2005.

[13] Dan Pelleg and Andrew Moore, " X-means: Estimating K-means with efficient estimation of the number of clusters.", International Conference on Machine Learning - ICML , pp. 727-734, 2000

[14] Li. W., & Henry, S., "Object-Oriented Metrics that Predict Maintainability", The Journal of Systems and Software, Vol. 23, pp. 111-122, 1993.

Author' biography with Photo

Astha Mehra is B. Tech (CS&E) student in Amity University Uttar Pradesh, India. Her research areas include Software Engineering and data mining techniques.

Sanjay Kumar Dubey is an Assistant Professor and Proctor in Amity University Uttar Pradesh, India. He has submitted his Ph. D. thesis in Object Oriented Software Engineering. He has published more than 73 papers in International Journals. He has presented 14 research papers at various National/International conferences. He is member of ACM, IET and IEANG. His research areas include Human Computer Interaction, Soft Computing, and Usability Engineering.