SOFTWARE FAULT INVESTIGATION USING K MEANS

Anantdeep kaur[1], Simerjeet kaur[2]
[1]Assistant Professor, [2]Research Scholar
[1,2]UCOE, Punjabi University, Patiala (Punjab), India

***ABSTRACT*: Software maintenance has been a necessity for the smooth running of the system. Large amount of data is available in different modes like videos, images etcetera. It has become very difficult to manage it time to time. Therefore, Cluster testing had been used in different ways to detect the faults in system. Moreover, Cluster testing works by making the clusters of data and perform testing that make it more effective and prompt. The current work is done to examine the effectiveness of different clustering algorithms such as k means algorithm, density based clustering, quad tree based clustering etcetera and to use the effective one to eliminate the problem of timely detection of errors in the software system**

***Keywords:* clustering, quad tree algorithm, k means clustering, software fault prediction, density based algorithm.**

I. INTRODUCTION

The procedure of clustering is basically to group the certain number of instances so that instances in the same group have similarity among themselves whereas dissimilar with other instances. There are various levels of keeping the maintenance phase of software up to date or we can say there are various techniques of analyzing software system for faults in it. However, one of them is cluster level approach that can be efficiently used to predict the phases or modules having errors. A cluster level analysis of software includes or considers the interactions among the cooperating classes groups. Cluster level testing is suitable for object oriented softwares. It is the grouping of collection of objects such that same group objects are alike in certain way as compared to other group elements. The main aim of clustering is to classify data according to the need of the user or as the data being instructed by the user to make it work to deliver desired results. There are different types of clustering that are explained below.

A. *Hierarchical clustering*

Connectivity based clustering, also known as *hierarchical clustering*, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

B. *Centroid based clustering*

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized types of clustering.

C. *Distribution based clustering*

The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution. While the theoretical foundation of these methods is excellent, they suffer from one key problem known as over fitting, unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

D. *Density based clustering*

In density-based clustering,[9] clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. The key drawback of DBSCAN and OPTICS is that they expect some kind of density drop to detect cluster borders. Moreover, they cannot detect intrinsic cluster structures which are prevalent in the majority of real life data. A variation of DBSCAN, EnDBSCAN,[14] efficiently detects such kinds of structures. On data sets with, for example, overlapping Gaussian distributions - a common use case in artificial data - the cluster borders produced by these algorithms will often look arbitrary, because the cluster density decreases continuously. On a data set consisting of mixtures of Gaussians, these algorithms are nearly always outperformed by methods such as EM clustering that are able to precisely model this kind of data

II. RELATED WORK

Swati M.Varade and Prof.M.D.Ingle, 2012 [1] gives the software architecture to forecast the software faults. They use hyper quad tree to find the centre of cluster for k-means algorithm that will find the error that tends to occur at the time of classification of the data. It depicts the efficiency of hyper

quad tree algorithm and proposes a system to efficiently anticipate the errors.

Supreet Kaur, and Dinesh Kumar, 2012 [2] demonstrate the efficiency of density based spatial technique of clustering and its evaluated to anticipate the faults in java based object oriented software. Weka tool is used in the evaluation to meet the requirements. Hence, it put forth the advantages of DBSCAN to predict the faults in software systems

Sarath et al. , 2012 [3] proposes different techniques to predict the faults in software systems. They used unsupervised learning approach to test the software for faults. The results of fault forecasting by quad tree based k-mean algorithm are comparable to other methods used for fault testing and are found better.

kaur et al. , 2012 [4] try to focus on timly finding of faults to avoid the inconvenience. In this different data mining techniques like association mining and classification, clustering are studied. This approach can be proven useful for the developers to fix errors or to check fault prone modules easily. It is also mentioned that unsupervised techniques can be used where there are no clues of faults to find the faults.

Sridhar et al. , 2012 [5] an intelligent traffic light that controls the flow of traffic was introduced. System detects the level of congestion and the abnormal situations in two main highways and four intersections. Collects the data and the information from a video imaging system by system, it is used to captures and interprets images, this image is used to detect and count the vehicle . This data will be sent to another system based on genetic algorithm. System used in the smart city platforms to manage the transportation system through the control of the intelligent traffic lights. The system depends on some theories and rules that made to order the priority and the green light interval time, genetic algorithm used in this proposed system and input to GA is based on the data that got from the video image detection

kanungo et al. , 2002 [6] put forth the implementation of k-mean algorithm. It requires the major data structure as a kd tree. They show the applications of filtering algorithm in two kinds. After the data sensitive analysis of the algorithm run time. It depicts that as the distance among clusters increases it runs fasts. Secondly, various empirical studies on synthetic data and real data sets from different applications presented. It demonstrates that the algo is very simple to implement and requires only the kd tree to be built for the data points.

Junjie Li et al. , 2008 [7] presents the implementation of the fuzzy k-mean algorithms for the numerical data so that the clustering process do not get sensitive to the initial cluster centre. In combination with cluster validation techniques new algorithm can infer the number of clusters in the data sets that the problem in k mean implementation. They have conducted experiments also that has given the positive results.

Srivastava et al. ,2005 [8] propose that by using differential geometric treatment of planar shapes there are various tools like agglomerative cluster of imaged object according to shape of the boundary of them. Another is probability model lear for cluster of shapes. Last is testing of new shapes under competing probability model. Clustering is performed using minimum variance type criteria with markov process also being used. We can deduce the shape by hypothesis testing and hierarchical testing

X.Wu et al. , 2008 [9] presents that k-mean the widely used algorithm for clustering. Its simplicity and effectiveness are remarkable among all the available methods [17].On the negative side, using -means to cluster high-dimensional data with, for example, billions of features is not simple and straightforward [18]; the curse of dimensionality makes the algorithm very slow. On top of that, noisy features often lead to over fitting, another undesirable effect. Therefore, reducing the feature selection, and optimizing the -means objective on the low-dimensional representation of the high-dimensional data is an attractive approach that not only will make -means faster, but also more robust [19], [20]

III. FUTURE WORK

A. Problem Formulation

The big problem is the software fault detection with the software. Most of the software systems are deployed having lots of defects in them that needed to be corrected for the ultimate functionality of the software system. So, in this way, the large number of efforts of the organization go in vain to reduce the faults of software products. In other words, they have to pay more attention towards the control of the quality and testing phase. There is no any good approach with the help of which we can find the fault prone prone modules. The ambition of metrics of software is to judge the quality of the system, that can be maintainability, defect density, fault proneness, understandability, and reusability etcetera.

The one way with which we can improve the task of identifying fault prone modules is with improvement in scheduling and controlling project. So, there is need to develop a system that is real time to solve this issue. As a result we have clustering approach that can be used for the detection of defect prone modules of the software to lessen the efforts spent on all the modules for finding errors. K mean clustering approach is applied. The data set and metrics are taken from the open source repository.

B. Objective

1. Study of the different clustering algorithms
2. Implementation of the k-mean algorithm to cluster the faulty and non faulty classes in Object Oriented software system.
3. The analysis of clustering algorithm for effective fault prediction in software system.

.

IV. CONCLUSION

The major key region of software fault is lack of proper inspection and timely detection of errors. The clustering is more suitable for managing large data sets for their efficiency as data is being getting huge day by day due lot of developments. So, it is proposed to use best clustering algorithm to eliminate the problem. .

REFERENCES

[1] Swati M.varade, Prof.M.D.Ingle, "Overview of Software Fault Prediction using Clustering Approaches and Tree Data Structure", The International Journal of Engineering And Science (IJES) , Volume 1, Issue 2, Pages- 239-242 , 2012, ISSN: 2319 – 1813, ISBN: 2319 – 1805.

[2] Supreet Kaur, and Dinesh Kumar, "Software Fault Prediction in Object Oriented Software Systems Using Density Based Clustering Approach", International Journal of Research in Engineering and Technology (IJRET) Vol. 1 No. 2 March, 2012 ISSN: 2277-4378.

[3] Partha Sarathi Bishnu and Vandana Bhattacherjee, "Software Fault Prediction Using Quad Tree-Based K-Means Clustering Algorithm", IEEE Transactions on knowledge and data engineering, Vol. 24, no. 6, June 2012.

[4] Ms. Puneet Jai Kaur, Ms. Pallavi, " Data Mining Techniques for Software Defect Prediction", International Journal of Software and Web Sciences 3(1), December, 2012-February, 2013, pp. 54-57, ISSN (Print): 2279-0063 ISSN (Online): 2279-0071.

[5]]. M. Bhanu Sridhar, Y. Srinivas, M. H. M. Krishna Prasad," Software Reuse in Cardiology Related Medical Database Using K-Means Clustering Technique", Journal of Software Engineering and Applications, 2012, 5, 682-686.

[6] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on pattern analysis and machine intelligence, Vol. 24, no. 7, july 2002.

[7] Mark Junjie Li, Michael K. Ng, Yiu-ming Cheung, Senior Member, IEEE, and Joshua Zhexue Huang, "Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters", IEEE transactions on knowledge and data engineering, Vol. 20, No. 11, November 2008.

[8] Anuj Srivastava, Member, IEEE, Shantanu H. Joshi, Washington Mio, and Xiuwen Liu, Member, IEEE, "Statistical Shape Analysis:Clustering, Learning, and Testing", IEEE Transactions on pattern analysis and machine intelligence, Vol. 27, No. 4, APRIL 2005.

[9] X.Wu et al., "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008