Role of Test Case Prioritization based on Regression Testing using Clustering

Yamini Pathania
Post Graduate Student
Dept. of Computer Science Engineering
Chandigarh University

Gurpreet Kaur
Assistant Professor
Dept. of Computer Science Engineering
Chandigarh University

ABSTRACT
Software testing is the most important part of software development life cycle. There are various types in software testing which have their own different functionalities. Among them regression testing is most useful functional type of testing which is done in the software maintence phase. This testing is used to check the errors when any change is made in the existing system. To make system efficient and effective, techniques of test case prioritization are use. The reduction in the cost of testing and fault detection capabilities of testing should be done by test case prioritization. This technique is also applied on different algorithms to improve their efficiency. Many clustering algorithms may also use test case prioritization method to increase the efficiency in code coverage. Prioritization techniques that incorporate a clustering approach and utilize code coverage, code complexity to increase the effectiveness of the prioritization.

Keywords
Software testing, Regression testing, Test case prioritization, Clustering, DBKmeans clustering.

1. INTRODUCTION
Software testing is the process of evaluating a system or its components with the intent to find that whether it satisfies the specified requirements or not and provide the stakeholders with information about the quality of the product or service under test. Testing in term of ANSI/IEEE 1059 standard may be defined as "A process of analyzing a software item to detect the differences between existing and required conditions (that is defects/errors/bugs) and to evaluate the features of the software item".

Testing mainly consists of two V's that are verification and validation. Verification is also known as static testing. In static testing, software are examine manually and missing requirements, design defects, syntax errors etc. are different types of defects find in static testing by review, inspection and walk through. Static testing can start early in the life cycle (e.g. by verifying User Requirement) and find bugs before compilation. Now, the term validation is also known as dynamic testing. In dynamic testing, it describes the dynamic behaviour of code by testing in which software executed by giving set off inputs, then examined it's output and finally compare the result what is expected. Dynamic testing can start after development of software component and find the bug after compilation and linking. Variables not constant, checking if from the expected values are different types of defects find in dynamic testing by unit testing, integration testing, and acceptance testing. Software testing is an important component of software quality assurance, and there are almost many software organizations which spending up to 40% of their resources on testing.

1.1.Types of Software Testing
There are two types of software testing, which is performed in software development life cycle.

1.1.1 *Manual Testing:* - In manual testing, tester takes over the role of an end user and test the software to identify any un-expected behaviour or bug. This kind of testing performed manually without using any automated tool or any script. To ensure the completeness of testing, tester use test plan, test cases or test scenarios to test the software.

1.1.2 *Automation testing:* - Automation testing is also known as test automation. In this type of testing, tester writes scripts and uses software to test the software. It is used to re-run the test scenarios that were performed manually, repeatedly and quickly. Automation testing mainly used in regression testing and apart from regression testing, it is also used to test the application from load performance and stress point of view. Automation testing increases the test coverage; improve accuracy, saves time as well as money in comparison to manual testing. This process involves automation of manual process.

1.2. Requirements to Automate System
- All GUI items, connections with databases, any change in the database, field validations etc. can be efficiently tested by automating the manual process or any area where large amount of users' can access the software simultaneously should be automated but we cannot automate everything in software.
- Test automation should be used when:

a) Requirements not changing frequently,
b) Projects that require testing the same areas frequently,
c) Stable software with respect to manual testing,
d) Availability of time,
e) Accessing the application for load and performance with many virtual users etc. are needed for software.

• Automation is done by using a supportive computer language like VB scripting and an automated software application. Many tools are available which can be used to write automation scripts like Selenium, IBM Rational Functional Tester, Testing Anywhere, WinRunner, LaodRunner, HP Quick Test Professional etc. The following process can be used to automate testing:

a) Identifying areas within a software for automation.
b) Selection of appropriate tool for Test automation.
c) Writing Test scripts.
d) Development of Test suits.
e) Execution of scripts.
f) Create result reports.
g) Identify any potential bug or performance issue.

2. REGRESSION TESTING

Regression testing is one of the most useful software testing types during software maintence. Regression testing may define as it is the testing process that is done to find the regression in the system after doing any changes in the product or it is used to check whether the new changes occurs the errors in the software or not. Program is tested before applying any change then again retested the program in the selected areas to detect whether the change created new bugs or issues, after a change is applied, or it may achieve its intended purpose after the actual change is made. This type of testing is essential for large organization. Now a day regression testing is handed via specialized testing tools and it is nearly impossible for human tester to perform the same tasks as crucial as automated software testers.

There is always one question arising in the software testers mind that why and when we perform the regression testing. Software maintenance is modification of product or software which includes deletion of existing features, enhancements, optimization and error corrections. Due to these modifications, it may cause the system to work incorrectly. For this reason, regression testing becomes necessary.

2.1. Techniques of Regression Testing

Three types of regression testing techniques are retest all, regression test selection and prioritization of test cases.

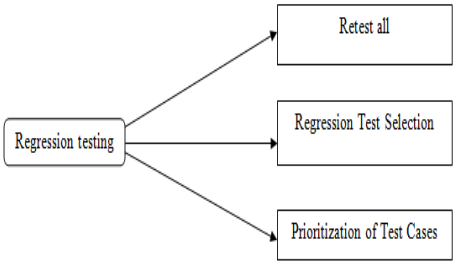

Fig.1 Techniques of regression testing

- Retest all is very expensive because it requires huge time and resource. In this method, the entire test suite in existing test bucket or suite should be re-executed.
- Regression test selection is better to select part of test suite to run rather then re- executing the all test suite. It may categorize in two ways: Reusable test cases which used in succeeding regression cycle and in obsolete test cases can't be used in succeeding cycle.
- Prioritization of test cases depending on business impact, and frequently used functionalities. In this method, selection of test cases based on priority will greatly reduced the regression test suite.
  The main goal of the test case prioritization schedule test cases in order to increase their ability to meet some performance goal: Rate of fault detection, Rate of code coverage and rate of increase of confidence in reliability. Rate of fault detection which a is a measure of how quickly the fault is detected so that during testing faster feedback can provide about system under testing and allow the software tester to correct the software at earlier phase as possible.

3. CLUSTURING

Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. There are various clustering techniques like k-means clustering, hierarchical clustering, DBSCAN clustering, OPTICS, STING etc.

3.1 K-means Clustering

In 1967 James MacQueen first proposed K- Means clustering algorithm. K-mean clustering is one of the simplest unsupervised learning algorithms and also one of the popular partitioning algorithms that solve the well known clustering problem. The idea is to classify the data into k clusters where k is the input parameter specified in advance through iterative relocation technique which converges to local minimum [18].

First phase is to determine k centers at random one for each cluster. Next phase is to determine distance between data points in Dataset and the cluster centers and assigning the data point to its nearest cluster. To calculate the distance generally Euclidean distance must be consider. Initial grouping is done when all the data points are included in some clusters. Then new centers are calculated by taking the average of points in the clusters. This is done because of inclusion of new points may lead to change in cluster centers. This process of centre updating is iterated until a situation where centers do not update anymore or criterion function becomes minimum.

3.1.1 Limitations:

a) K- means has a problems when cluster are of differing
- Size
- Density
- Non- globular shape

b) K- means has a problem when data contains outliers.

3.2 DBSCAN Clustering

DBSCAN stands for Density Based Spatial Clustering and Application with Noise**.** It is one of the most common clustering algorithms. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. It depends upon density reachability and density connectivity. Density reachability defines whether two distance close points belong to the same cluster.

Density connectivity is the last building step of DBSCAN. Points p0 and pn are density connected, if there is a sequence of density reachable points p1,i2,...,i(n-1) from p0 to pn such that p(i+1) is density reachable from pi [6].

3.2.1 Limitations

a) DBSCAN can only result in a good clustering as good as its distance measure is in the function get Neighbors (P, epsilon). The most common distance metric used is the Euclidean distance measure. This distance metric can be rendered almost useless especially for high dimensional data.

b) DBSCAN does not respond well to data sets with varying densities (called hierarchical data sets).

3.3 DBKMEANS CLUSTERING

Clustering is the process of partitioning a group of data points into a small number of clusters. The objective of clustering is to partition a set of objects into clusters such that objects within a group are more similar to one another than patterns in different clusters.

DBKmeans clustering algorithm is a combination of DBSCAN and K-mean clustering. This algorithm performs better then DBSCAN when handling clusters of circularly distributed data points and slightly overlapped clusters. The criteria for splitting or joining a cluster can be decided based on the number of expected points in a cluster or the expected density of the cluster (derived by using the number of points in a cluster and the area of the cluster) [6]. There is lot of scope for the DBKmeans clustering algorithm in different application areas such as medical image segmentation and medical data mining. Basically DBKmeans clustering algorithm overcome the drawbacks of DBSCAN and K-means clustering algorithms.

4. LITERATURE REVIEW

R.Krishnamoorthi and S.A.Sahaaya Arul Mary [5] propose a new test case prioritization technique using Genetic Algorithm (GA). The proposed technique prioritizes subsequences of the original test suite so that the new suite will have a superior rate of fault detection when compared to rates of randomly prioritized test suites which is run within a time-constrained execution environment.

K.Mumtaz, and Dr K. Duraiswamy, [6] a novel density based k-means clustering algorithm has been proposed to overcome the drawbacks of DBSCAN and k-means clustering algorithms and the result provide an improved version of k-means clustering algorithm.Arvinder Kaur, and Shubhra Goyal [7] describe a new Genetic Algorithm to prioritize the regression test suite is introduced that will prioritize test cases on the basis of complete code coverage.

Carlson, Ryan, Hyunsook Do, and Anne Denton [8], represent new prioritization techniques that incorporate a clustering approach and for real faults, utilize code coverage, code complexity, and history data. This paper describes a clustering approach to help improve test case prioritization. Arvind Kumar Upadhyay, A. K. Misra [11] propose clustering based prioritization and support their effort with average percentage of fault detection (APFD) measure.

Md. Junaid Arafeen and Hyunsook Do [12] describe requirements-based clustering approach that incorporates traditional code analysis information can improve the effectiveness of test case prioritization techniques. Muthusamy, Thillaikarasi [16] describes a proposed algorithm which is based on analysis of the percentage of test cases performed to find the faults and on APFD metric's results. To improve the regression testing process, test case prioritization techniques used to prepare the execution level of test cases. Manika Tyagi and Sona Malhotra, [17] describes a 3-phase approach to solve test case prioritization. Firstly, we are removing redundant test cases by simple matrix operation. In next phase, test cases are selected from the test suite which reduces execution time and minimal sets cover all faults and in last phase, we assign priority to test cases obtained from the second phase.

Medhun Hashini D.R [20] describe test case prioritization that utilizes a clustering approach can improve the effectiveness of test case prioritization techniques. Goyal, M., and S. Kumar[18] proposed Mean-Based algorithm is easy to implement and it proves to be a better method to determine the initial centroids which can be used in the k-means clustering algorithm and give better clusters equally for uniform and non-uniform data sets.

5. CONCLUSION

The cost of software development is directly dependent on the testing effort. To reduce the cost by prioritizing test cases and running the tests for the selective test cases as per the available time and manpower. Test case prioritization involves scheduling of test cases in an order that increases their effectiveness in meeting some performance goals. One such goal is APFD(average percentage of faults detected) measure that increases the chances of finding faults earlier in the software testing lifecycle and may facilitate the ultimate goal of software development by improving quality. This technique is applied on many algorithm such as genetic algorithm, PSO, MOPSO, ant colony optimization etc. We can apply the same test case prioritization technique for the proposed DBKmeans clustering to increase its efficiency and also increase the fault detection rate. This is also helpful to improve test case prioritization using this algorithm.

6. REFERENCES

[1] Elbaum, Sebastian, Alexey G. Malishevsky, and Gregg Rothermel. "Test case prioritization: A family of empirical studies", IEEE Transactions on Software Engineering, 2002.

[2] Ng, Raymond T., and Jiawei Han, "Clarans: A method for clustering objects for spatial data mining", IEEE Transactions on Knowledge and Data Engineering, 2002.

[3] Srivastava, Praveen Ranjan, "Test case prioritization", Journal of Theoretical and Applied Information Technology, 2008.

[4] Liela Naslavsky, Hadar Ziv, Debra J. Richardson, "A Model based Regression Test Selection Technique", International Conference on Software Maintenance (ICSM), IEEE, 2009.

[5] Krishnamoorthi, R., and SA Sahaaya Arul Mary, "Regression test suite prioritization using genetic algorithms", International Journal of Hybrid Information Technology, 2009.

[6] K. Mumtaz, and Dr K. Duraiswamy. "A novel density based improved k-means clustering algorithm– Dbkmeans", International Journal on computer science and Engineering, 2010.

[7] Arvinder Kaur and Shubhra Goyal. "A genetic algorithm for regression test case prioritization using code coverage" International journal on computer science and engineering, 2011.

[8] Carlson, Ryan, Hyunsook Do, and Anne Denton, "A clustering approach to improving test case prioritization: An industrial case study", 27th IEEE International Conference on Software Maintenance (ICSM), 2011.

[9] J. Wang, Y. Zhuang, C. Jianyun, "Test Case Prioritization Technique based on Genetic Algorithm", International Conference on internet Computing and information Services, Hong Kong , 2011.

[10] Huang, Yu-Chi, Kuan-Li Peng, and Chin-Yu Huang, "A history-based cost-cognizant test case prioritization technique in regression testing", Journal of Systems and Software, 2012.

[11] Upadhyay, Arvind Kumar, and A. K. Misra. "Prioritizing Test Suites Using Clustering Approach in Software Testing.", International Journal of Soft Computing and Engineering (IJSCE), 2012

[12] Arafeen, Md Junaid and Hyunsook Do, "Test case prioritization using requirements-based clustering", IEEE Sixth International Conference on Software Testing, Verification and Validation (ICST), IEEE, 2013.

[13] Kumar, Mukesh. "An optimized farthest first clustering algorithm", Nirma University International Conference on Engineering (NUiCONE), IEEE, 2013.

[14] P Kandil, S Moussa, N Badr,"Regression testing approach for large scale", IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), 2014.

[15] Manika Tyagi and Sona Malhotra, "Test case prioritization using multi objective particle swarm optimizer", 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT), IEEE, 2014.

[16] Muthusamy, Thillaikarasi, "A New Effective Test Case Prioritization for Regression Testing based on Prioritization Algorithm", International Journal of Applied, 2014.

[17] Tyagi, M.; Malhotra, S. "Test case prioritization using multi objective particle swarm optimizer", International Conference on Signal Propagation and Computer Technology (ICSPCT), 2014.

[18] Goyal, M., and S. Kumar, "Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability." Journal of The Institution of Engineers (India), 2014.

[19] Hettiarachchi, Charitha, Hyunsook Do, and Byoungju Choi, "Effective Regression Testing Using Requirements and Risks", 2014 Eighth International Conference on Software Security and Reliability, IEEE, 2014.

[20] Medhun Hashini D.R, "Clustering approach to test case prioritization using code coverage metric" International Journal of Engineering and Computer Science, 2014.

[21] P Kandil, S Moussa, N Badr,"Regression testing approach for large scale", IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), 2014.