International Journal of Information Technology and Management

Vol. IX, Issue No. XIII, August-2015, ISSN 2249-4510

TEST SUITE REDUCTION USING DATA MINING

www.ignited.in

International Journal of Information Technology and Management
Vol. IX, Issue No. XIII, August-2015, ISSN 2249-4510

Test Suite Reduction Using Data Mining

Arjun Singh

B.Tech, IV Year, Northern India, Engineering College, Delhi, India

Abstract – Software testing is a process used to identify the correctness, completeness, and quality of developed computer software. It includes a set of activities conducted with the intent of finding errors in software so that it could be corrected before the product is released to the end users. The practical methods commonly used to detect the presence of errors in a program are to test it for a set of inputs called test case. A Test Case is a set of actions executed to verify a particular feature or functionality of software application. When designing the test case, the redundant test cases are formed that are of no use, increases the testing effort and increase the cost and time of testing. In this paper, the goal is to reduce the time spent in testing by reducing the number of test cases. For this the data mining approach of clustering technique is used in software testing to reduce the test suite. Mining of test case will improve the efficiency of software testing.

Keywords : Software Testing, Data Mining, Clustering, Test Suite Reduction

- - - - - - - - - - - - - - X - - - - - - - - - - - - - -

1. INTRODUCTION

Testing software is a very important and challenging activity. Nearly half of the software production development cost is spent on testing. The main objective of software testing is to eliminate as many errors as possible to ensure that the tested software meets an acceptable level of quality. The tests have to be performed within budgetary and scheduled limitations. An important activity in testing is test case design. Many programming groups are relying more and more on automated testing, it requires a well-developed test suite of testing scripts in order to be truly useful. If the Test suites tend to grow in size as software evolves, then testing becomes too cost to execute entire test suites. The test suite reduction techniques significantly reduce the size of the test suites. In this paper, the application of data mining techniques with software testing is used for reducing the size of the test suite (Reliable Mining of Automatically Generated Test Cases from Software Requirements Specification, 2010). The less, the number of test cases, the time taken for executing the program should also be less. This consequently improves the effectiveness of the test process (The Data Mining Approach to Automated Software Testing), (SE Code Optimization using Data Mining Approach, 2012).

Software engineering is the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software, and the study of these approaches; that is, the application of engineering to software. A part of Software Engineering is to do Software Testing which consists of a set of activities of verifying and validating that a software application or program meets the business and technical requirements that guided its design and development and works as expected and also identifies important errors or flaws categorized as per the severity level in the application that must be fixed. (Lilly Raamesh et al., 2009).

Essence of software testing is to choose a representative value (known as test case) from the input to perform the programs under test. The actual results of the programs will be checked to verify the consistency with the expected ones. If the results are different, it should take some correction, adjustment and evaluation correspondingly. (Kartheek Muthyala et al.)

Among the different approaches we may distinguish between specification-oriented approaches (or black-box testing), which generate the test cases from the program specification, and implementation-oriented approaches (or white-box testing), which generate the test cases from the code of the program under test. Test cases have to be generated according to the test adequacy criterion, which is considered to be a stopping rule that determines whether sufficient testing has been done and provides measurements of test quality (Amalgamation of Automated Testing and Data Mining, 2011).

A test suite is a collection of test cases for particular software. Redundancy of test cases will be possible in the software. Redundancy is the repetition of data, between one test case and the other. So it is obvious that the reasonable structure of test suite is one of the key points in software testing achieve by which lot of time can be saved from executing redundant or unnecessary test cases (Lilly Raamesh et. al., 2010), (Extracting Test Cases by Using Data Mining, 2011). This replicated data isn't visible enough to capture unless and until the sophisticated techniques like data mining is used. In this paper, the proposed methodology of data mining technique is used with software testing to remove the redundant test case so that the test suites are reduced or minimized (Mining Test Cases, 2012)..

2. REVIEW OF LITERATURE

Test case prioritization techniques are playing a vital role in testing (Harrold 2001) as researches have shown that half of the total cost of software development consists of testing activities. Research in regression testing spans a wide range of topics.

In testing, usually the program is tested with a set of test data known as test cases to uncover the errors. The same test cases may be used for running a program again after changes in the program due to some cause like change in requirement, change in technology etc., is known as regression testing. It has been tested experimentally (Coley 2007) (Cost-Constrained Data Acquisition for intelligent Data Preparation, 2005). that some of the biggest causes for project failures are lack of user input and changing or incomplete requirements. Software engineers save the test cases and re-run the test cases as regression test in later versions. (Prioritizing Test Suites Using Clustering Approach in Software Testing, 2012).

Different environments can assist regression testing particularly automation of test case execution in the regression-testing phase. The main disadvantage of regression testing is that the additional cost, time, manpower etc. that are needed for testing the program again for defects. But the additional cost, time, manpower etc., can be reduced to some extent by not executing the full set of test cases again. So the techniques like test case reduction, test case optimization and test case prioritization (Bates and Horwitz 1993, Binkley 1995, Elbaum et al 2000, Malishevsky et al 2006, Rothermel 1999, Tonella 2006, Walcott 2006, Yau and Kishimoto 1987) may be used. (Prioritizing Test Suites Using Clustering Approach in Software Testing, 2012), (UML Generated Test Case Mining Using ISA, 2011), (An Efficient Algorithm for Reducing, 2013).

Numerous prioritization techniques have been described in the research literature (Elbaum et al 2001,2002, Jones and Harrold, 2001, Rothermel et al 2001 and Wong et al 1997). Studies (Elbaum et al 2002, Rothermel et al 1999, 2001) have shown that at least some of these techniques can significantly increase the rate of fault detection of test suites in comparison to the rates achieved by unordered or randomly ordered test suites. These early indications of potential are encouraging, however, studies have also shown that the rates of fault detection produced by prioritization techniques can vary significantly with several factors related to program attributes, change attributes, and test suite characteristics (Elbaum et al 2001, 2003). (A Comparative Study of White Box, 2012).

In several instances, techniques have not performed as expected. In the empirical studies (Elbaum et al 2002), however, it was often observed that the results are contrary to this expectation. It is possible that engineers choosing to prioritize for both coverage and change attributes may actually achieve poorer rates of fault detection than if they prioritized just for coverage, or did not prioritize at all.

Feature selection is considered as the most essential step of many pattern recognition and artificial intelligence problems (Zhang 2007). In feature selection the mutual information (Shannon 1948) is acting as standard measure of dependence which is used for feature selection and ranking as a filter in many fields like medicine, neuroscience, genomics and related fields, ecology, economics, etc (Ding and Peng 2005, Kwak and Choi 2002, Peng et al 2005). Mutual Information has been perfectly utilized in the approach called mRMR (minimum Redundancy-Maximum Relevance) (Peng et al 2005) which aims at obtaining maximum classification or prediction performance with a minimal subset of variables by reducing the redundancies among the selected variables to a minimum and to maximize their relevance. (Different Approaches to White Box Testing Technique for Finding Errors, 2011).

In feature selection, an approach was introduced for building efficient classifiers using weak features **(**Günter and bunke 2004, Oliveira et al 2006 , Drauschke and forstner 2008) from a group of classifiers. Feature selection can also be used for handwritten script recognition (Oliveira et al 2003, Günter and bunke 2004, Morita et al 2003). (Control Flow graphs And Code Coverage, 2010).

Optimization is also used in many clustering methods like dynamic clustering (wang 2006), fuzzy clustering (zhi et al 2004), traveling salesman problems (Xu and Xiao 2006) , knapsack problem (Goldbarg 2006), and minimum spanning trees (Zhu 2006). Combinatorial optimization was also used in path optimization (wang 2006), vehicle routing (wa 2004, Cernic 1999).

Ant colony optimization (ACO) (Yogesh Singh et al 2010) is used in regression testing. For document clustering analysis, (Xiaohui Cui et al 2006) flocking based approach was proposed. The proposed flock-clustering algorithm utilized the stochastic and heuristic principles for monitoring bird flocks or fish schools.

The Hybrid Particle Swarm Optimization (HPSO) algorithm (Arvinder Kaur et al 2011) was used for performing an efficient regression testing. Here, the HPSO is a combination of Particle Swarm Optimization (PSO) method and Genetic Algorithms (GA), to extend the search space for the solution. (Research on the Application of Data Mining in Software Testing and Defect Analysis, 2009).

3. MINING TECHNIQUES FOR TEST SUITE REDUCTION

Data mining is the process of extracting patterns from data. As more data are gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

While data mining can be used to uncover patterns in data samples, it is important to be aware that the use of non-representative samples of data may produce results that are not indicative of the domain. Similarly, data mining will not find patterns that may be present in the domain, if those patterns are not present in the sample being "mined". There is a tendency for insufficiently knowledgeable "consumers" of the results to attribute "magical abilities" to data mining, treating the technique as a sort of all-seeing crystal ball. Like any other tool, it only functions in conjunction with the appropriate raw material: in this case, indicative and representative data that the user must first collect. Further, the discovery of a particular pattern in a particular set of data does not necessarily mean that pattern is representative of the whole population from which that data was drawn. Hence, an important part of the process is the verification and validation of patterns on other samples of data.

The term data mining has also been used in a related but negative sense, to mean the deliberate searching for apparent but not necessarily representative patterns in large numbers of data. To avoid confusion with the other sense, the terms *data dredging* and *data snooping* are often used. Note, however, that dredging and snooping can be (and sometimes are) used as exploratory tools when developing and clarifying hypotheses. (Defect Data Analysis Based on Extended Association Rule Mining, 2007).

4. APPLYING DATA MINING CONCEPTS

There are many methods available for mining different kinds of data, including association rule, characterization, classification, clustering, etc.

We can utilize any of these techniques based on What kind of data bases to work on What kind of knowledge to be mined What kind of techniques to be utilized We can apply association or clustering techniques for test case mining.

Association

Association rules describe the association among items in the large database. For example, one may find, from a large set of transaction data, such as association rule as if customer buys (one brand of) milk, he/ she usually buys (another brand of) bread in the same transaction. Using these association rules, we can derive the association patterns from large databases.

Data classification

Data classification is the process, which finds the common properties among a set of objects in a database and classifies them into different classes, according to a classification model.

Clustering

Clustering is the process of grouping the data into classes or clusters so that object within a cluster has high similarity in comparison to another, but is dissimilar to object in other clusters. It doesn’t require the class label information about the data set because it is inherently a data driven approach. It is the process of grouping or abstract object into classes of similar object.

Among all the mining techniques, clustering is the most effective technique, which we are going to use for test case mining.

Clustering analysis helps constant meaningful partitioning of a large set of object based on a “divide and conquer” methodology, which decomposes a large-scale system into smaller components to simplify design and implementation. As a data mining task, data clustering identifies cluster or densely populated regions, according to some distance measurement, in a large, multidimensional data. Given a large set of multidimensional data points, the data space is usually not uniformly occupied by the data points. Data clustering identifies the sparse and the crowded places, and hence discovers the overall distributions patterns of the data set.

For cluster analysis to work efficiently and effectively as many literatures have presented, there are the following typical requirements of clustering in data mining.

- Scalability:
- Ability to deal with different types of attributes: o Discovery of clusters with arbitrary shape:

- Minimal requirements for domain knowledge to determine input parameters:

CONCLUSION

In this paper, a new approach to automatically generate test cases from SRS and mining of test cases has been discussed. Firstly a formal transformation of a detailed SRS to a UML state model, secondly the generation of test cases from the state model and lastly mining of Test cases. The introduction of agents can bring enhancement.

REFERENCES

A Comparative Study of White Box (2012). Black Box and Grey Box Testing Techniques, International Journal of Advanced Computer Science and Applications, Vol. 3, No.6.

Amalgamation of Automated Testing and Data Mining (2011). A Novel Approach in Software Testing, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2.

An Efficient Algorithm for Reducing the Test Cases which is Used for Performing Regression Testing (2013). 2nd International Conference on Computational Techniques and Artificial Intelligence (ICCTAI' 2013) March 17-18.

Automatic Software Test case Generation (October 2012). An Analytical Classification Framework, International journal of Software Engineering and its Applications, Vol. 6, No. 4.

Control Flow graphs And Code Coverage (2010). Int. J. Appl. Math. Comput. Sci., Vol. 20, No. 4, pp. 739–749.

Cost-Constrained Data Acquisition for intelligent Data Preparation, (2005). IEEE Transactions on Knowledge and Data Engineering, Vol 17, No 11.

Defect Data Analysis Based on Extended Association Rule Mining (2007). Shuji Moriaski, Arito Monden, Tomoko Matsumura, Fourth International Workshop On Mining Software Repositories, IEEE.

Different Approaches to White Box Testing Technique for Finding Errors (2011). International Journal of Software Engineering and Its Applications Vol. 5 No. 3.

Extracting Test Cases by Using Data Mining (2011) Reducing the Cost of Testing, International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150- 7988 Volume 3, pp. 730-737.

Kartheek Muthyala et al., A Novel Approach To Test Suite Reduction Using Data Mining, Indian Journal of Computer Science and Engineering (IJCSE).

Lilly Raamesh et al., (2009). Knowledge Mining of Test Case System, International Journal on Computer Science and Engineering Vol.2 (1), 69-73

Lilly Raamesh et. al., (2010). An Efficient Reduction Method for Test Cases, International Journal of Engineering Science and Technology, Vol. 2(11), 6611-6616.

Mining Test Cases (2012). Optimization Possibilities, International Journal on Advances in Software, vol. 5 no 3 & 4, http://www.iariajournals.org/software/.

Prioritizing Test Suites Using Clustering Approach in Software Testing (2012). International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-4.

Reliable Mining of Automatically Generated Test Cases from Software Requirements Specification, (2010). International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 3.

Research on the Application of Data Mining in Software Testing and Defect Analysis (2009). Yanguang Shen, jie Liu, IEEE Second International Conference on Intelligent Computation Technology and Automation.

SE Code Optimization using Data Mining Approach, (2012). International Journal of Computer & Organization Trends –Volume 2, Issue 3.

The Data Mining Approach to Automated Software Testing.

UML Generated Test Case Mining Using ISA (2011). International Conference on Machine Learning and Computing, IPCSIT vol.3.