Research on Application of Software Testing Based on Data Mining

Hengyao Tang [1], Xiaoyan Zhan [2]

[1] Computer School of Huanggang Normal University, Huanggang 438000, China

[2]School of Foreign Studies Huanggang Normal University, Huanggang 438000, China

Abstract

This paper briefly introduces the classification of software testing and the basic realization process of data mining technology. It focuses on the application of several methods of data mining technology in software test case generation and test result analysis. It has a certain guiding significance for software testing work.

Keywords

Data mining, software test, Software defects.

1. Introduction

With the rapid development of the software industry, the scale and complexity of the software have increased significantly. In order to improve the quality of the software, the software testing as an important part of the software development process and the whole process of the software life cycle has been paid more and more attention. After years of development, software testing has been relatively the formation of a relatively stable test process, test methods, designed a variety of software testing model. Software testing technology also put forward some new test technology with the emergence of new computer technology. This paper discusses the application of data mining technology in software testing.

2. Software Testing Classification

Software testing refers to the use of test cases by manual or automatic means to run or test a system process, the purpose is to find errors, test whether to meet user needs or to find out the expected results and actual results of the difference. At present, the software testing methods and techniques are more, each method and technology has its own characteristics, different test purposes and the environment can use different test methods and techniques.

Software testing can be divided into static testing and dynamic testing from the point of view of the need to perform the test software. If you do not actually run the software under test, but only static check code, interface or document may exist in the error, known as static test; dynamic test is the actual operation of the test program, enter the appropriate test data, check the actual output of the results And whether the expected results are consistent with the process.

From the perspective of whether the software internal structure and specific implementation, software testing can be divided into white box test, black box test, gray box test. White box testing, also known as structural testing, transparent box testing, logic-driven testing or code-based testing, its focus on the internal logic of the program structure, all the logical path to test; black box test focus on the external structure of the program, regardless of internal logic structure, Mainly for the software interface and software functions to test, only to check whether the program in accordance with the requirements of the specifications of the normal implementation of the specification.

From the perspective of the software development process, software testing can be divided into unit testing, integration testing, validation testing, system testing, acceptance testing, regression testing. Unit testing, also known as module testing, is the smallest unit (program module) for software design, the correctness of the test work, the purpose is to find the various modules may exist within the various errors; integration test is a number of software units integrated Whether the software meets the requirements of the specific intended use, verifies that the test verifies that the software meets the

requirements specified in the software requirements specification; the system test is done by verifying the test software , As an element of the entire computer-based system, with the computer hardware, peripherals, some support software, data and personnel and other system elements together, in the actual operating environment, the computer system for a series of assembly testing and confirmation Testing; acceptance test is user-oriented test, by the user to participate in the design of test cases, the use of the actual production of data to test whether the software system to meet the requirements of the two sides agreed; regression test is modified after the old code, carry out testing Confirm the modification does not introduce new bugs or cause other code to generate an error.

3. The Application of Data Mining Technology in Software Testing

3.1 Data mining technology

Data mining technology is a process of extracting useful knowledge from large data sets (which may be incomplete, including noise, ambiguous, uncertain, and even based on various forms of storage) Even the massive data found in the implied meaningful knowledge. It is usually divided into three steps: information collection, data integration, data specification, data cleanup, data transformation, data mining process, pattern evaluation and knowledge representation. It can be used to classify data, such as classification, estimation, forecasting, association rules and clustering.

3.2 Test case generation based on data mining

A very important job in software testing is to enter a large number of test data execution programs to check if the actual output matches the expected results. And the input data of the program is usually an infinite set, the software test data is limited. How to quickly and accurately find a limited set of test data to test the infinite input data may exist problems, software testing technology is very concerned about the problem. Designing test cases only by equivalent class division and boundary value analysis methods It is difficult to test the implicit procedural defects that exist in a complex and highly correlated program. The use of intelligent algorithms such as genetic algorithm and ant colony algorithm in data mining technology to extract useful test cases from input data has become a very important means in software testing.

Genetic algorithm to generate test cases [1] is: from the beginning of the initial population generation, according to the characteristics of the measured module to determine the fitness function, the implementation of the program, the assessment of each test case of the fitness, the higher the fitness test case and The closer the effect is expected, and then use the three basic operators to improve the test case by using crossover, mutation, and evaluate with the fitness function until the best expectation can end the process.

Ant colony algorithm to generate test cases is to put a group of ants in the path of the map, through the ant movement, release pheromone and other behavior search the optimal path to participate in the path of the ants, on the basis of knowledge, the application of ants The group algorithm completes the knowledge reasoning and generates the required test data.

Decision Tree Selection Test Case Method [2] is: First, the decision tree begins with a single node representing the training sample. If the samples are in the same class, the node becomes a leaf and is marked with that class. Otherwise, with a measure as a heuristic information, select the best classification of the attributes of the sample, as the node's test properties. For each known value of the test attribute, create a branch and divide the sample accordingly. Repeat the above process, the recursive formation of each partition on the sample sub-decision tree. When all samples of a given node belong to the same class or have no remaining attributes to further divide the sample or when there is no sample in the branch, the recursion is stopped, then the majority of the voting is used to convert the given node to the leaf and the majority of the parent node Class to mark.

Analysis of Test Results Based on Data Mining

3.3.1 Sequence pattern mining

Sequence pattern mining was first proposed by Agrawal and Srikant: Given a set of different sequences, where each sequence is ordered in a sequence of different elements, each element is composed of different items, given a user-specified Minimum support threshold, the task of sequence pattern mining is to find all the frequent subsequence, that is, the frequency of not less than the minimum support threshold of the sub-sequence [3].

In an event-driven application, the code does not execute according to a predetermined path but executes a different code snippet when responding to a different event. The order of the events determines the order in which the code is executed, the disorder of the events, and the ordering of the code segments. Due to the different implementation of the order, may lead to the same code in the execution of the context of the different, resulting in potential errors. Tests for event-driven programs are more difficult, especially for complex programs where the sequence of possible event combinations is almost unlimited.

Event sequences are generally cluttered, contain large amounts of redundant data, but events that cause software errors occur in the sequence, and their type and order are usually constant. We can use the sequence pattern mining to analyze the sequence of events in the program, from a number of sequence of events to dig out more frequent common sub-sequence, the use of the sub-sequence to reproduce the error, and then locate the error, to avoid a large number of error reports one by one analysis.

3.3.2 Association rule analysis

In the software development process, there will be some software defects between the relevant, such as the type of software in the definition of error defects, often the corresponding data overflow overflow. An association rule is a reflection of the interdependence or association between an event and other events. In the software test, you can introduce the association rule analysis technology, analyze the error data found in a large number of historical tests, find out the possible error type association rules, and provide support for the follow-up software testing to find potential errors that have not yet been found. For example, in the historical test data, there is a $\wedge$ b $\Rightarrow$ c association rule between the defect type a, the defect type b and the defect type c, and the defect type a and the defect type b have been found in the current software test but not found Defect type b, then need to further test to confirm the existence of defect type c[4].

3.3.3 Cluster analysis

Clustering is the process of classifying data into several classes or clusters. Objects in the same cluster have very similarity, and the objects between different clusters have great dissimilarity. The goal of clustering analysis is to collect data on a similar basis to classify, and to perform more accurate classification learning under the condition that the data distribution is unknown.

In the software test can be found in the test error or defect results data set, based on test methods, error types, test phase, development (testing) personnel and other indicators of cluster analysis, found that the distribution of various types of data. Through the distribution of the situation, found that the developer error rate was significantly higher, the test method to detect more types of errors, the testers are good at finding software errors and so on. This can be in the subsequent test work, you can find the focus of different products testing, but also for different products for comparison, assessment, targeted to develop strategies to optimize the testing process. For example, a developer often often make mistakes in the definition of type, in the test will need to focus on the design of its design module type of problem; in the black box test found that more defects occurred in the implementation of a certain type of query function, then Need to classify the query function as a class, according to its characteristics to design a special test case [5].

3.3.4 Cluster analysis

Classification analysis is different from clustering analysis, which is a supervised learning that knows in advance the training sample label, whose input set is a set of records and several given markers, and records are sorted by mark. The purpose of classification is to find the exact description and model for

each class by recording the manifestation. Classification widely used methods are decision trees, neural networks and radial basis functions. In the software test using classification analysis method, according to the test results show defect attributes for defect type classification is a very important work.

4. Conclusion

Data mining technology has provided new ideas and methods for software testing. This paper studies the application of representative data mining technology in software testing in recent years, and summarizes the application of several data mining methods in software test case generation and test result analysis, and explains its use principle. For the software to adopt multi-channel testing, reduce the immune function of software defects, to provide theoretical help.

Acknowledgments

This work is supported by the Excellent Youth Project of Hubei Provincial Department of Education (No. Q20132904) and the project of Huanggang Normal University (No. 2014016603).

References

[1] Wang Hao,Xie Junkai,Gao Zhongyi:Genetic Algorithm and Its Application in Software Test Data Generation, Computer Engineering and Applications,(2001), p.64-68.
[2] Ma Jing, Gu Jingwen:Application of decision tree in software test case generation,Computer Technology and Development,(2008), p.66-69 .
[3] Jia Ning:Research on Software Testing Technology Based on Data Mining (MS.Tianjin University School of management, China 2007).
[4] Liu Jie:Research on Software Defect Based on Positive and Negative Association Rules (MS. Hebei Engineering University, China 2010).
[5] Chen Yuan:Research on Software Defect Prediction Technology Based on Data Mining(MS. Chinese Academy of Sciences, China 2012).