An Empirical Model of Fault Prediction in Software Engineering
Karri Santoshkumar*, Tammineni Ravikumar
1Final Year M.TechStudent, Dept. of CSE, Aditya Institute of Technology and Management(AITAM), Tekkali, Srikakulam, Andhra Pradesh
2Sr Asst. Professor, Dept. of CSE, Aditya Institute of Technology and Management(AITAM), Tekkali, Srikakulam, Andhra Pradesh

Abstract: Even though various approaches developed by the many researchers, they may not be optimal while predication of faults. In this approach we are introducing the fault prediction approach with OO metrics along with cyclomatic complexity and nested block depth, in acceptance testing, each function specified in the design document can be independently tested, that is, a set of test cases is developed for each function, not for each workflow module or other module/component. Our experimental results show the efficient fault prediction with our computation parameters. Our approach mainly concentrates on the count of faults prior to testing, expected number of faults, our classification which involves algorithmic and processing, control, logic and sequence, typographical syntax errors i.e. incorrect spelling of a variable name, regular iteration of statements, incorrect initialization statements per module, this proposed classification approach shows optimal results while analyzing the metrics with training samples after measurement.

Index Terms: Metrics, Cyclomatic complexity, OO metrics, Fault prediction

I. INTRODUCTION
Software quality always depends on various factors like complexity, quality & size and it is not an end mark at customer satisfaction, even though customer satisfaction is the ultimate goal, quality of product is always basic feature of software development. Main developing software products, reusability is the basic important feature to use the existing code, it reduces the redundancy and complexity while execution of client request [1].

Traditional testing process is time consuming and expensive while handling of large projects [2]. The traditional approaches are fault prediction works with the basic metrics like Lines of code(LOC). Number of errors found, Number of errors found with respect to the module, These parameters are not sufficient to measure the fault prediction and cost effectiveness. Software faults may be design faults which are deterministic in nature and are identified easily and other type of software faults is classified as being temporary internal faults that are transient and are difficult to be detected through testing[3].

It is difficult to analyze the fault prediction by simply measuring the software metrics of the project. We require a classification tool for the analysis of the predicted results[4]. The use of software in high-assurance and mission-critical systems increases the need to develop and quantify measures of software quality. Subsequently, software metrics are useful in the timely prediction of high-risk components during the software development process such a prediction enables software managers to target quality improvement efforts to the needed areas. For example, prior to the system test, identifying the components that are likely to be faulty during operations can improve the effectiveness of testing efforts. Various software quality modeling techniques have been developed and used in real-life software quality predictions.

Classification-based modeling for software quality estimation is a proven technique in achieving better software quality. Some classification techniques used for software quality estimation include optimal set reduction, logistic regression, decision trees and neural networks, and case-based reasoning. Software metrics-based quality classification models classify software modules into groups [5].

A two-group classification model is the most commonly used, in which program modules are grouped into classes such as, fault-prone (fp) and not fault-prone (nfp). Software analysts interested in predicting whether a software component will be fp or nfp, can utilize classification models to obtain the needed prediction. When associating with such models, two types of misclassification errors are encountered: Type I error, that is, an fp component is predicted as fnp, and Type II error, i.e., an fnp component is misclassified as fp. Practically speaking, Type II errors are more severe in terms of software development costs and the organization’s reputation. They may involve inspection and correction to components after they have been deployed for operations [6].

In our previous empirical studies related to software quality classification modeling, we have investigated several classification techniques, including classification and regression trees (CART), tree-based classification with S-PLUS, the Tree disc algorithm, the C4.5 algorithm, the Sprint-Sql algorithm, logistic regression, and case-based reasoning. Classification models
were calibrated using case studies of different large scale software systems, including the one presented in this paper. The models are calibrated to classify software modules as either \( f_p \) or \( n_f_p \), as defined by the system being modeled.

**II. RELATED WORK**

The target organization is a software purchaser-side company that provides various types of telecommunication services using acquired software systems. In the software acquisition processes, the company is responsible for requirements analysis, architectural design, and acceptance testing, while developer-side companies are in charge of detailed design, programming unit/integration/ system testing, and debugging. As the services grow in the number of variations with shorter renewal cycles than ever before, the main motivation here is optimization of acceptance testing to provide high quality services to customers. From this perspective, the primary goal of this paper is reduction of acceptance test effort using techniques for predicting fault-prone modules [7].

Our study includes metrics collection, building predictor models, and assessing the reduction of test effort.

**C4.5 Algorithm**

The C4.5 algorithm is an inductive supervised learning system which employs decision trees to represent a quality model. C4.5 is a descendent of another induction program, ID3, and it consists of four principal programs: decision tree generator, production rule generator, decision tree interpreter, and production rule interpreter. The algorithm uses these four programs when constructing and evaluating classification tree models. Different tree models were built with varying parameters: minimum node size before splitting and pruning percentage.

The C4.5 algorithm commands certain preprocessing of data in order for it to build decision tree models. Some of these include attribute value description type, predefined discrete classes, and sufficient number of observations for supervised learning. The classification tree is initially empty and the algorithm begins adding decision and leaf nodes, starting with the root node.

**Tree disc Algorithm**

The Tree disc algorithm is a SAS macro implementation of the modified CH_{2}-square Automatic Interaction Detection algorithm. It constructs a regression tree from an input data set that predicts a specified categorical response variable based on one or more predictors. The predictor variable is selected to be the variable that is most significantly associated with the dependent variable according to a chi-squared test of independence in the contingency table.

Regression tree-based models are built by varying model parameters in order to achieve the preferred balance between the misclassification error rates, and to avoid over fitting of classification trees. A generalized classification rule is used to label each leaf node after the regression tree is built. This classification rule is very similar to the approach followed, when using S-PLUS regression trees as classification trees [8].

**Sprint-Sliq algorithm**

Sprint-Sliq is an abbreviated version of Scalable parallelizable induction of decision Trees-Supervised Learning In Quest, the algorithm can be used to build classification tree models that can analyze both numeric and categorical attributes. It is a modified version of the classification tree algorithm of CART, and uses a different pruning technique based on the minimum description length principle. The algorithm has excellent scalability and analysis speed. Classification tree modeling using Sprint-Sliq is accomplished in two phases: a tree building phase and a tree pruning phase. The building phase recursively partitions the training data until each partition is either “pure” or meets the stop-splitting rules set by the user. The IBM Intelligent Data Miner tool, which implements the Sprint-Sliq algorithm, was used by our research group to build classification trees. Sprint-Sliq uses the Gini Index to evaluate the goodness of split of all the possible splits. A class assignment rule is needed to classify modules as \( f_p \) and \( n_f_p \) [9].

**Logistic Regression**

Logistic regression is a statistical modeling technique that offers good model interpretation. Independent variables in logistic regression may be categorical, discrete or continuous. However, the categorical variables need to be encoded (e.g., 0, 1) to facilitate classification modeling. Our research group has used logistic regression to build software quality classification models.

Let \( x_i \) be the \( j_m \) independent variable, and let \( x_i \) be the vector of the \( i_m \) module’s independent variable values. A module being \( f_p \) is designated as an “event”. Let \( q \) be the probability of an event, and thus \( q = q / (1 - q) \) is the odds of an event. The logistic regression model has the form [10],

\[
\log(q / (1 - q)) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m
\]

Where, \( \log \) means the natural logarithm, \( b_j \) is the regression coefficient associated with independent variable \( x_j \), and \( m \) is the number of independent variables. Logistic regression suits software quality modeling because most software engineering measures do have a monotonic relationship with faults that is inherent in the underlying processes. Given a list of candidate independent variables and a significance level, \( a \), some of the estimated coefficients may not be significantly different from zero. Such variables should not be included in the final model.
Architecture

III. PROPOSED WORK

We are proposing an efficient fault prediction with efficient parameters or metrics optimal results. The paper is concerned with faults due to Object oriented issues and measures with the integrated basic metrics of software engineering. Our proposed approach measures the metrics on the source code and analyzes the Testing samples which are measured Algorithmic and processing, Control, logic and sequence, Typographical Syntax error etc, then forwards these metric measures to the classification rule for analysis of testing sample, the below diagram shows proposed architecture, initially complexity, cyclomatic and object oriented metrics applied over source code blocks and measures respective parameters and treat them as testing samples and forward those testing samples to training samples to compute posterior probability of testing sample.

Basic Complexity metrics

Basic entity for measuring the software quality or cost are the metrics we initially computing the complexity metrics as Hallstead’s metric, which includes program length, program vocabulary which leads to the computation of Estimated length and finally computes the purity ration, then computes the program effort in terms of volume and density.

\[
\begin{align*}
n_1 &= \text{the number of distinct operators} \\
n_2 &= \text{the number of distinct operands} \\
N_1 &= \text{the total number of operators} \\
N_2 &= \text{the total number of operands}
\end{align*}
\]

From these numbers, several measures can be calculated:

- Program vocabulary: \( n = n_1 + n_2 \)
- Program length: \( N = N_1 + N_2 \)
- Calculated program length:
  \[
  N' = \eta_1 \log_2 \eta_1 + \eta_2 \log_2 \eta_2
  \]
- Volume: \( V = N + \log_2 \eta \)
- Difficulty: \( D = \frac{n_1}{2} \times \frac{N_2}{\eta_2} \)
- Effort: \( E = D \times V \)

Cyclomatic complexity

It is a quantitative measure of logical strength of the program. It directly measures the number of linearly independent paths through a program's source code, these paths can be represented in various formats as sequence, while for iteration, if then else for decision making, until for iteration and then computes the complexity in terms of...
vertices or nodes and edges and in terms of number of predicate nodes. Mathematically, the cyclomatic complexity of a structured program is defined with reference to the control flow graph of the program, a directed graph containing the basic blocks of the program, with an edge between two basic blocks if control may pass from the first to the second. The complexity \( M \) is then defined as
\[
M = E - N + 2P,
\]
where
\[
E = \text{the number of edges of the graph.}
\]
\[
N = \text{the number of nodes of the graph.}
\]
\[
P = \text{the number of connected components.}
\]

Object oriented metrics

Object oriented metrics are also one type of metrics for measure the quality and cost of the project in terms of
- Weighted methods per class
- Depth of the inheritance tree
- Coupling between classes
- Response for a class
- Class size and finally computes the polymorphism factor which includes the factors of number of overriding methods, number of methods and number of sub classes

Classification

After the computation of all the metrics, testing sample data can be forwarded to the training samples to measure and classify the metric results with existing results in terms of conditional probability by using Naive Bayesian classifier, which leads to the fault prediction analysis.

Naive Bayesian Classification

Estimating probabilities

\( P(X), P(X|C_i), \text{and } P(C_i) \) may be estimated from given data

Bayes Theorem:
\[
P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}
\]

Steps Involved:
1. Each data sample is of the type
   \( X=(x_i) \ i = 1(1)n, \text{where } x_i \text{ is the values of } X \text{ for attribute } A_i \)
2. Suppose there are \( m \) classes \( C_i, i = 1(1)m \).

\( P(C_i|X) > P(C_j|X) \)
i.e BC assigns \( X \) to class \( C_i \) having highest posterior probability conditioned on \( X \)

The class for which \( P(C_i|X) \) is maximized is called the maximum posterior hypothesis.

From Bayes Theorem

3. \( P(X) \) is constant. Only \( P(X|C_i), P(C_i) \) need be maximized.
   - If class prior probabilities not known, then assume all classes to be equally likely
   - Otherwise maximize \( P(C_i) = S/S \)

Problem: computing \( P(X|C_i) \) is unfeasible!

4. Naive assumption: attribute independence

\[
P(X|C_i) = P(x_1,\ldots,x_n|C_i) = \prod P(x_i|C_i)
\]

5. In order to classify an unknown sample \( X \), evaluate for each class \( C_i \). Sample \( X \) is assigned to the class \( C_i \) if

\[
P(X|C_i)P(C_i) > P(X|C_j)P(C_j).
\]

Experimental Analysis

We experimentally analyzed the results by computing the all measures over code snippets individually and then computed parameters forwarded towards training dataset.

Let us consider an example for computing Halstead metrics for small code snippet as follows

```c
main()
{
    int a,b,c,avg;
    scanf("%d %d%d", &a,&b,&c);
    avg=(a+b+c)/3;
    printf("avg=%d",avg);
}
```

1. \( n_1=10, n_2=7, n_3=17 \)
2. \( N_1=15, N_2=31 \)
3. Calculated Program Length \( N'=10 \times \log_2 10 + 7 \times \log_2 7=52.9 \)
4. Volume \( V=31 \times \log_2 17=126.7 \)
• Difficulty D=10/2 \times 15/7=10.7
• Effort E=10.7 \times 126.7=1,355.7
• Time required to program T=1,355.7/3000=0.004

For the measure of object oriented metrics, we considered weighted methods per class, Depth of the inheritance tree, Coupling between classes and overridden methods and for the measure of cyclomatic complexity we considering an example like

if A = 354
then if B > C
then A = B
else A= C
endif
endif
print A

Thus using the formal formula the cyclomatic complexity is 8-7 + 2 = 3. In this case there is no graph called or subroutine. Alternatively one may calculate the cyclomatic complexity using the decision points rule. Since there are two decision points, the cyclomatic complexity is 2 + 1 = 3. Now these measures can be forwarded to training dataset for classification of this testing sample, the following experimental results shows efficient acceptance results than the traditional approaches as follows

Fig 2: Comparative Analysis

IV. CONCLUSION

We are concluding our research work with efficient metrics computation with Complexity, cyclomatic, object oriented metrics over blocks of source code followed by naive Bayesian classification to compute whether the testing sample is cost effective or not by analyzing fault prediction

REFERENCES

[2] Predicting Defect Densities in Source Code Files with Decision Tree Learners Patrick Knab, Martin Pinzger, Abraham Bernstein
[3] Comparative Assessment of Software Quality Classification Techniques: An Empirical Case Study TAGHI M. KHOSHOFTAAR, NAEEM SELIYA

BIOGRAPHIES

Karri Santosh kumar completed B.Tech degree in Computer Science and Engineering from Sri Venkateswara College of Engineering and Technology, Etcherla. He is pursuing M.Tech degree in the Department of Computer Science and Engineering, from Aditya Institute of Technology And Management (AITAM), Tekkali, A.P, and India. His Interested areas are Software Engineering and Data Mining.

Tammineni Ravi Kumar received the B.tech in CSIT and M.Tech in CSE degrees from the department of Computer Science and Engineering at the Jawaharlal Nehru Technological University, Hyderabad(JNTUH) respectively. He is currently working as Sr. Asst. Prof in Department of Computer Science &Engineering, in Aditya Institute of Technology and Management, Tekkali, Andhra Pradesh, India. He has 7 years of experience in teaching Computer Science and Engineering related subjects. His Research interests include Software Engineering, Data Mining and Software Testing Methodologies.