Volume 3, Issue 6, June 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Novel Methodology for Software Reliability Using Mixture Models and Non-Homogeneous Poisson Process

Y. Vamsidhar*
Department of CSE
Swarnandra College of Engg, &Tech, Narsapur, India

Y. Srinivas**
Department of IT
GIT, GITAM University, Visakhapatnam, India

A. Brahmani Devi***
Department of CSE
GIT, GITAM University, Visakhapatnam, India

Abstract-Software Reliability addresses the problem of predicting the failure rate of the software under development. These predictions help to model the software so that the output software is free from faults. Many of the Software Reliability models presented in literature are based on the identification of the fault either in the testing phase or while debugging the software. Software Reliability process in this paper is highlighted by using a mixture model together with a Non Homogenous Poison Process model. The reliability of proposed model is evaluated by using Least Square Estimation Method and Goel Model.

Keywords-Software Reliability, Mixture Models, Failure Data, Defects, Non-Homogeneous Poisson Process (NHHP), Least Square Estimation Method, Goel Model.

I. Introduction

Software Reliability Growth Models aim at the identification of faults and defects while developing software or manufacturing a product, which in turn results in the betterment of the products. Many Models have been proposed by the researches about software reliability based on Non-Homogeneous Poisson process (NHPP) [1] [2] [3] [4] [5].The main reason for selecting the NHPP technique, is it facilitates the testers an analytical framework which helps in identifying the faults derived from the software during the testing process. However NHHP models will be more effective if the errors are identified or predicted exactly [6] [7]. But in the software testing phase there is a possibility that the errors may not be identified exactly which may lead towards the misclassification of errors. The misclassification may be due to the network failure, transmission delay and other factors, the faults are assumed to be known before hand, but in practical, the faults can be only identified only after testing the software for long duration, i.e. the faults in the software are considered to follow a unsupervised learning mechanism. Hence in order to predict these faults and categorize these faults appropriately. in the proposed model, we have considered a new symmetric Mixture models distribution to model the faults appropriately.Inorder to demonstrate the proposed work we have considered database from NASA Metrics Data Pros(MDP) which consists of two modules: KC1, KC2. KC_2 contains over 3000 modules written in C++ program. Out of these modules 414 modules are identified to contain faults and each module contains a minimum of 37 Lines of code (LOC) and one module containing a maximum 1275 LOC. The KC_2 dataset consists of 3 attributes; error rate, defect and defect density.Inorder to identify the exactness of the faults, we have considered a module containing 2011 LOC.Inorder to identify the faults exactly the median of the failure dataset is obtained and the fault values are categorized into groups by the condition that if the value is greater than median value, it is attributed as fault and assigned binary value 1 if the value is less than median value it is assigned to 0.Thus the data which is clustered into 2 groups as fault and Non-fault, are given to the new symmetric distribution and Probability Density Function (PDF) for each of the data in both groups are obtained. These PDF values are considered for further classification. The main advantage of using new symmetric distribution is that basing on their dataset; we can model either a symmetric curve or Non-symmetric curve. It consists of a shape parameter which considers two values 0 and 1,if the value is 0 it behaves a simple Gaussian Mixture Model and if the value is greater than 1 it behaves as a new symmetric distribution. The remaining paper is organized as follows. The Section-II of paper deals with New Symmetric Mixture Model Distribution, Section-III of paper deals with Model Selection Criteria. Section-IV of paper deals with Methodology, Section-V of paper deals with Non-Homogeneous Poison Paper (NHHP) and results derived are concluded in Section-VI.

II. New Symmetric Mixture Model Distribution

The failure dataset, generated may be symmetric/asymmetric. The New Symmetric Mixture Model is used to cater both Symmetric and Asymmetric Distribution. The Probability Density Function (PDF) of the new symmetric Distribution.

$$f(x) = \left[\frac{2r + [(\frac{x-\mu}{\sigma})^2]^r e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma(2r)^r(2\pi)^{\frac{1}{2}} + \sum_{i=1}^{r} \binom{r}{i} (2r)^{r-i} 2^{i+\frac{1}{2}} \Gamma\left(i+\frac{1}{2}\right)\sigma} \right]$$

Where r= shape parameter, μ-mean, Γ=gamma function, $\sigma = standard\ deviation$

The shape parameter r take two values 0 and 1, if r=0 it behaves as a Gaussian distribution and if r=1 it behaves as New Symmetric Distribution.

III. Model Selection Criteria

While classifying the data into several classes which contain similar characteristics one need to have prior knowledge to divide the data into classes. Generally in software development process, we come across two classes: fault prone and non-fault prone[8][9]. In the traditional approaches of software Reliability models these classes are classified without prior knowledge which makes the system supervised. However the faults can be estimated after testing for long time and it follows a unsupervised mechanism which needs to be assessed. For these purpose models like Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are used to identify the classes in unsupervised data. Among these models AIC model is used for our purpose since it estimates the unidentified data more exactly [10].

The AIC model is given by

$$AIC\ (k) = 2k - 2\log(L)$$

Where L=log likelihood estimate, k= no. of parameters.

IV. Methodology

The data is considered KC_2 NASA dataset which is tabulated in Table: 1 is considered and from the dataset two different classes fault prone zones and Non-fault zones are identified. By estimating the median the initial categorization of classes is carried out. The data after getting categorized into groups are model using AIC criteria, based on Log likelihood estimate obtained from the PDF of New Symmetric Distribution, presented in Section-II.The dataset is given to the model which categorizes the data and the faults are estimated. These model acts as a classifier to validate the no. of faults in the developed model. Metrics proposed by Goel Okumoto (1979) is utilized.

V. Non-Homogeneous Poisson Process(Nhpp)

The estimated errors are model further using the software reliability growth model to Goel-Okumoto model is utilized.

The classification of fault and non-fault is done based on the heuristic of Liu et al. So, based on the median value the dataset is classified into two classes and it is tabulated in Table 2.

From the histogram of the data, it can be visualized that it forms a bell shaped distribution and hence, r, the shape parameter is assumed to be 0 which formulates a Gaussian distribution. The formula for Gaussian distribution is:

$$G = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\left(\frac{x-M}{\sigma}\right)^2}$$

Where M=median value,$\sigma = standard\ deviation$,

x =errors in a dataset

TABLE: I

KC_2 NASA DATASET

S. No.	Defects	Cumulative of defects
1	75	75
2	81	156
3	86	242
4	90	332
5	93	425
6	96	521
7	98	619
8	99	718
9	100	818
10	100	918
11	100	1018

The Median for the above dataset is:

M=96 and N=11

TABLE: II

CLASSIFICATION OF FAULT AND NON-FAULT

Fault(>96)	Non-fault(<=96)
98	75
99	81
100	86
100	90
100	93
	96

The probability Density Function for each error is calculated and tabulated as below:

TABLE: III

PDF VALUE USING GAUSSIAN DISTRIBUTION

Defects	PDF Value
75	0.00466
81	0.0048
86	0.00492
90	0.00497
93	0.00499
96	0.0050
98	0.004996
99	0.004992
100	0.00498
100	0.00498
100	0.00498

These PDF values calculated for each for each error in a dataset, the Log likelihood function by generated by considering the LOG (PDF) value. And we calculate the AIC by using a formula, proposed in Section-III of the paper.
The AIC values are tabulated below:

TABLE: IV

AKAIKE INFORMATION CRITERIA (AIC) VALUES

SL No	Fault	log likelihood	AIC
1	75	-2.33	6.6
2	81	-2.31	8.6
3	86	-2.30	10.6
4	90	-2.303	12.6
5	93	-2.301	14.6
6	96	-2.3010	16.6
7	98	-2.3013	18.6
8	99	-2.3017	20.6
9	100	-2.302	22.6
10	100	-2.302	24.6
11	100	-2.302	26.6

Based on the AIC values we calculate the mean value function by using Goel-Okumoto model. The mean value of function is:

$$m(t) = a(1 - e^{-bt})$$

Where t=time in seconds
a, b are the parameters
m(t)= no. of expected errors at time t
The parameters of Goel-Okumoto Model are estimated by using Least Square estimation method. The values are:

A=94.9

B=2.4

Therefore $m(t) = 94.9(1 - e^{-2.4(1)})$

m(t)=86.35

The mean value function of Goel is applied on fault data which is obtained from the heuristic of Liu et al.S

TABLE: V
FAULT DATASET FROM CLASSIFICATION OF FAULT AND NON-FAULT

Fault(>96)
98
99
100
100
100

The values of a, b are obtained by using Least Square estimation method.
a=99.4
b=0.5
Therefore $m(t) = 99.9(1 - e^{-0.5(1)})$
m(t) =39.11
The expected failure rate in this model decreases when compare to model based on Goel.

VI. Conclusion

This paper addresses software Reliability Growth Model based on New Symmetric Distribution. The main advantage of this model is that it identifies the faults in a unsupervised manner which is more desirable in software applications where the faults can only be attributed after regress testing. The model developed is tested on a KC_2 database and results derived are compared to that of the mean value function of the model proposed by Goel. The developed model exhibits m(t) value which indirectly assumes that this model outperforms the existing models and can be used for real-time application of generating a software.

References

[1] Yamada,S.,H. Ohtera and H. Narihisa, 1986.*Software Reliability Growth Models with testing* - effort IEEE Trans. Reliability,35:19-23.
[2] Goel,A.L. and K.Okumoto,1979.*Time Dependent Error-Detection Rate Model for Software Reliability and other performance measures.Transactions Reliability*, 28:206-211.
[3] Hossain,S.A. and R.C. Dahiya,1993.*Estimating the parameters of a Non-Homogeneous Poisson Process Model for Software Reliability*. IEEE Transactons Reliability, 42:604-612.
[4] Littlewood,B.,1981.Stochastic Reliability-Growth: *A Model for fault-removal in computer programs and between design*.IEEE Trans. Reliability 30:313-320.
[5] Musa,J.D.,A.Iannino and K Okumoto, 1987.*Software Reliability: Measurement, Predicyion and Application*. McGraw-Hill, NewYork.
[6] Chang-Hua Hu et al (2010) *System Reliability Prediction Model based on Evidential reasoning algorithm with non linear optimization,Exper Systems with applications*,2010,PP 2550-2562.
[7] S.M.K Quadn et al (2011) *Software Reliability Growth Model with Generalized Experimental tests-Effort and Optimal Software Release Policy*, Global Journal of CS & Teon,Vol,11 (12) 2011,PP 26-43
[8] Zhang, X., Teng, X. And Pham ,H. *Considering Fault Removal Efficiency In Software Reliability Assessment ,IEEE Transactions on Systems*, Man and Cybernetics-part A,Vol.33,No.1,2003;114-120.
[9] Liu et al (2010) Software Reliability growth Model selection and Combination: *A New Approach, Computer Engineering and Applications*.
[10] K.S.Rao et al 1997,*On New Symmetric Distribution*, Journal of Indian Society of Agriculture and Statistics,Vol. 50 (11),1997,PP : 95-102.

