ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

A Survey on Software Defect Prediction Using Data Mining Techniques

Varsha G. Palatse, Prof. V. S. Nandedkar

ME Student, Dept. of Computer Engineering P.V.P.I.T, Bavdhan, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering P.V.P.I.T, Bavdhan, Pune, Maharashtra, India

ABSTRACT: Software testing plays a vital role in software development especially when the software developed is mission, safety and business critical applications. Software testing is the most time consuming and costly phase. Prediction of a modules info fault-prone and non fault prone prior to testing is one of the cost effective technique. Predicting a safe module as faulty increases the cost of projects by more cautious and better-test resources allocation for those modules, whereas prediction of faulty code as fault free code end up in under-preparation and may leave modules untested this may cause accidental failure and lead towards massive loss . In this research, we present a novel fault prediction technique that reduces the probability of false alarm (pf) and increases the precision for detection of faulty modules. The general expectation from a predictor is to get very high probability of false alarm (pf) to get more reliable and quality software product. Software Reliability is becoming an essential attribute of any software system. It is a significant factor in software quality since it quantifies software failures. Software defect prediction models have gained considerable importance in achieving high software reliability. Software defect prediction model helps in early detection of faults and contribute to their efficient removal and producing a reliable software system. This paper presents the survey on existing data mining techniques used for prediction of software defects.

KEYWORDS: Data Mining; Software Defect Prediction; Software Reliability

I. INTRODUCTION

The quality of software components should be tracked continuously during the development of high-assurance systems such as telecommunication infrastructures, medical devices, and avionic systems. Quality assurance group can improve the product quality by allocating necessary budget and human resources to low quality modules identified with different quality estimation models. Software defect, defined as deviation from expectation of software operation that might lead to software failures or any imperfection related to software itself, leading to huge economic loss, is an important issue in software development life cycle[2]. As Software development is a human activity, a lot of defects may be generated during the software development life cycle. It is quite difficult to develop fault free, quality software because of increasing complexity and the constraints under which software is developed. Defective Software poses considerable risk by increasing the development and maintenance costs and customer dissatisfaction. Moreover, software development companies cannot risk their business by providing defective low quality software.

It is, therefore of great concern to locate fault prone software modules at an early stage of the project. Tracking the fault as early as possible in software development process will not only improve the effective cost but also helps to achieve customer satisfaction and reliability of software developed[4]. Developing reliable, fault free and high quality software system is a complex and expensive task. It is beneficial to predict the faults because it helps in estimating test effort, reducing cost and developing a high quality and reliable software. Software defect prediction is the process of finding defective modules in the software[7].

Software Defect Prediction Model refers to those models that try to predict potential software defects from test data. There exists a correlation between the software metrics and the fault proneness of the software. A Software defect prediction models consists of independent variables (Software metrics) collected and measured during software development life cycle and dependent variable (faulty or non faulty). Firstly, the model is developed using the training data i.e. independent and dependent variables of previously developed Software. Then this model can be used to predict the defect of software in future[6].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

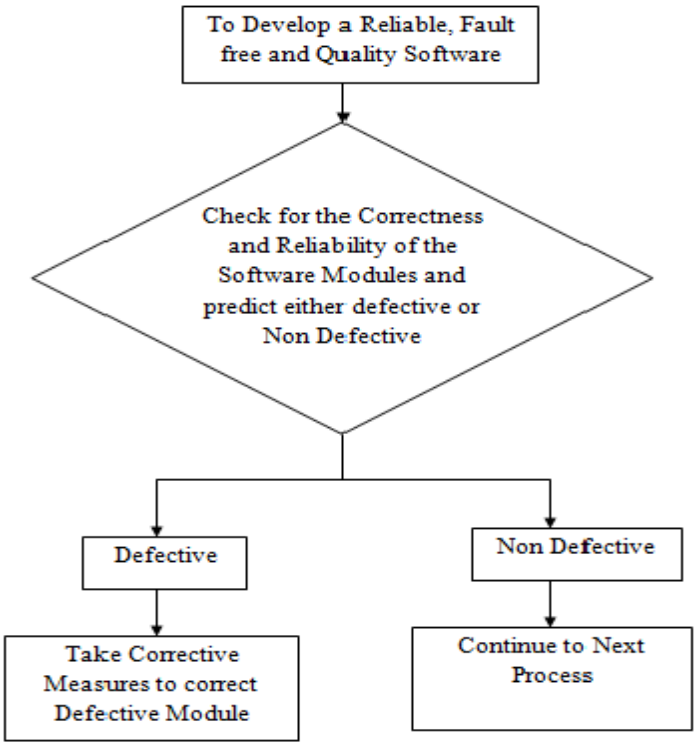

Figure 1: Software Defect Prediction

II. RELATED WORK

The growing complexities of software and increasing demand of reliable software have led to the progress of continual research in the areas of effective software reliability assessment. Software defect prediction is not a new thing in software engineering domain. A number of Software Defect Prediction models and techniques have been proposed by different researchers in recent years. In this section, some important contributions in this area are presented.

Wangshu Liu, Shulong Liu, Qing Gu, *Member, IEEE*, Jiaqiang Chen, Xiang Chen, *of* Empirical Studies of a Two-Stage Data Preprocessing Approach for Software Fault Prediction[1] provide a two-stage data preprocessing approach, which incorporates both feature selection and instance reduction, to improve the quality of software datasets used by classification models for software fault prediction. In the feature selection stage, we propose a novel algorithm which involves both relevance analysis and redundancy control. In the instance reduction stage, we apply random under-sampling to keep the balance between the faulty and non-faulty instances. We systematically design experiments based on the Eclipse and NASA datasets, and compare our approach to other commonly used data preprocessing methods.

Karunanithi et.al[2] of Software Defect Prediction Based on Classification Techniques presented the neural network model for software reliability prediction and found that neural network models are better at endpoint prediction than analytical models. They used different networks like feed forward NN, Jordan NN, recurrent neural networks.

Yajnaseni Dash, Sanjay Kumar Dubey, "Quality Prediction in Object Oreinted System[3], Software Defect Prediction Based on Clustering Techniques authors proposed a novel software defect prediction method based on functional clusters of programs to improve the performance. Until then, most methods proposed in this direction predict defects by class or file. Experiments carried out concluded that cluster based models can significantly improve the recall from 31.6 % to 99.2% and precision from 73.8 % to 91.6%. k-means based clustering approach has been used for finding the fault proneness of the Object oriented systems and found that k-means based clustering techniques shows 62.4%

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

accuracy. It also showed high value of probability of detection and low value of probability of false alarms. This study confirms the feasibility and usefulness of k means based software fault prediction models.

Qinbao Song, Martin Shepperd, Michelle Cartwright, Carolyn Mair, "Software Defect Association Mining and Defect Correction Effort Prediction [4], researchers proposed prediction of defect association and defect correction method based on association rule mining methods. The proposed methods were applied to defect data consisting of more than 200 projects over 15 years. It was concluded from experimental results that accuracy achieved is high for both defect association prediction and defect correction prediction. The results obtained were also compared with PART, C4.5 and Naive Bayes method and showed the accuracy improvement by 23 percent.

Gabriela Czibula, Zsuzsanna Marian, Istvan Gergely Czibula, "Software defect prediction using relational association rule mining", Information Sciences [5] , researchers proposed a novel defect prediction model based on relational association rules which are an extension of ordinal association rules and describe numerical orderings between attributes that commonly occur over dataset. This proposed model was evaluated on open source datasets and compared to similar existing approaches and found that this model over performed for most of the existing machine learning based techniques for defect prediction.

Kriti Purswani, Pankaj Dalal, Dr. Avinash Panwar, Kushagra Dashora, "Software Fault Prediction using Fuzzy C-Means Clustering and Feed Forward Neural Network In the paper[6], a hybrid approach based on K-Means Clustering and feed forward neural network has been proposed and it was found that performance is better in case of this hybrid approach as compared with the existing approaches in terms of accuracy , mean absolute error and root mean square error values.

Yasutaka Kamei, Akito Monden, Shuji Morisaki, Ken-ichi Matsumoto, "A Hybrid faulty module Prediction using Association Rule Mining and Logistic Regression Analysis [7] proposed Hybrid fault prone module prediction method was introduced that combines association rule mining with logistic regression analysis. If a module satisfies the premise of one of the selected rules, the module is classified by rule as either fault prone or not. Otherwise, the module is classified by the logistic regression. The prediction performance of this model was evaluated and compared with three other fault prone modules based on logistic regression model, linear discriminant model and classification tree. The experimental results showed improvement in performance as compared to conventional methods.

Lan Guo, Yan Ma, Bojan Cukic, Harshinder Singh, "Robust Prediction of fault proneness by Random Forest", Software Reliability Engineering In [8], researchers present a methodology for predicting software faults based on random forest, which is an extension of decision tree learning. Random forest technique was applied in five case studies based on NASA data set. The predictive accuracy of this technique was found to generally higher than that of achieved logistic regression, discriminant analysis and the algorithms in two machine learning software Packages WEKA .

III. CONCLUSION

After study of various researches related to data mining techniques for software defect prediction, we got that data mining is an emerging approach for defect prediction. Machine Learning Classifiers have emerged as a way to predict the fault in the software system. Since most of these studies have been performed using different data sets, reflecting different software development environment and processes, it is difficult to conclude the best software prediction model. Various models and techniques are studied which have their associated merits and demerits. The objective of this study is to analyse the performance of various data mining techniques used in software defect prediction models.

REFERENCES

[1] Wangshu Liu, Shulong Liu, Qing Gu, *Member, IEEE*, Jiaqiang Chen, Xiang Chen, *Member, IEEE*, and Daoxu Chen, *Member, IEEE* , "Empirical Studies of a Two-Stage Data Preprocessing Approach for Software Fault Prediction." in IEEE TRANSACTIONS ON RELIABILITY Volume: PP Year: 2015.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

[2] N.Karunanithi, D.Whitley, Y.K. Malaiya, "Using Neural networks in software reliability prediction" , IEEE Software,Vol.9, no.4, pp. 53-59,1992..
[3] Yajnaseni Dash, Sanjay Kumar Dubey, "Quality Prediction in Object Oreinted System by Using ANN: A Brief Survey" , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 2, Issue 2 , February 2012..
[4] Qinbao Song, Martin Shepperd, Michelle Cartwright, Carolyn Mair, "Software Defect Association Mining and Defect Correction Effort Prediction", IEEE Transaction on Software Engineering Vol.32, No. 2, February 2006, pp 69-82.
[5] Gabriela Czibula, Zsuzsanna Marian, Istvan Gergely Czibula, "Software defect prediction using relational association rule mining", Information Sciences, Volume 264 pages 260-278, April 2014.
[6] Kriti Purswani, Pankaj Dalal, Dr. Avinash Panwar, Kushagra Dashora, "Software Fault Prediction using Fuzzy C-Means Clustering and Feed Forward Neural Network" International Journal of Digital Application & Contemporary Research Volume 2, Issue 1 , July 2013.
[7] Yasutaka Kamei, Akito Monden, Shuji Morisaki, Ken-ichi Matsumoto, "A Hybrid faulty module Prediction using Association Rule Mining and Logistic Regression Analysis", Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and measurement, Pages 279-281.
[8] Lan Guo, Yan Ma, Bojan Cukic, Harshinder Singh, "Robust Prediction of fault proneness by Random Forest", Software Reliability Engineering, 2004 ISSRE 2004.

BIOGRAPHY

Varsha G. Palatse student of ME Computer Engineering second year from the college TSSM's Padmabhushan Vasantdada Patil Institute of Technology, Bavdhan, Pune.

Prof. V. S. Nandedkar is a faculty in the Computer Engineering from the college TSSM's Padmabhushan Vasantdada Patil Institute of Technology, Bavdhan, Pune, India.