A Survey of Different Software Fault Prediction Using Data Mining Techniques Methods

Karpagavadivu.K[1], Maragatham.T[2], Dr.Karthik.S[3]

Abstract- **Software fault prediction method used to improve the quality of software. Defective module leads to decrease the customer satisfaction and improve cost. Software fault prediction technique implies a good investment in better design in future systems to avoid building an error prone modules. Faulty modules are predicted using data mining techniques. Various classifiers are used to classify faulty or non-faulty modules. The goal of software fault prediction is to help the software manager to increase the software system quality. There are many software fault prediction techniques are available. This paper presents the survey on software fault prediction models.**

Key Words- **Clustering, Fault prediction, Machine learning, Quad tree, Software metrics, Software Quality.**

I. INTRODUCTION

Data mining is one of the evolution techniques in information technology. It can be named as “knowledge mining from data”. Before storing data into data warehouse or any type of databases, there is important to perform some data preprocessing steps. The preprocessing steps are data cleaning, integration, selection, transformation, pattern evaluation and knowledge presentation [1]. Data mining includes forecasting what may happen in future, classifying things into groups by recognizing patterns, clustering things into groups based on their attributes and associating what events are likely to occur together. Data mining process is reliable process and repeatable process by the people with small quantity of data mining skills.

Software quality exists wherever quality is defined in a business context. Software quality may be structural or functional quality. Structural quality refers how it meets nonfunctional requirements that support delivery of functional requirements. Functional quality shows how well it implies with software quality is characterized by some attributes like reliability, usability, efficiency and portability [ISO 01]. Software fault affects the software quality, so software fault prediction is an important one. Software quality prediction is performed at a time of software development life cycle and makes the efficient use of resources. The software quality prediction is performed by identifying the prediction of module is faulty or non-faulty.

Fault-prone prediction models are efficient and accurate. Fault-proneness models are built from information about the code and its faults [2]. Software metrics are used as independent variables and fault data are regarded as dependent variable in software fault prediction models [3]. Software metrics represent quantitative description of program attributes and the critical role play in predicting the quality of software [4].

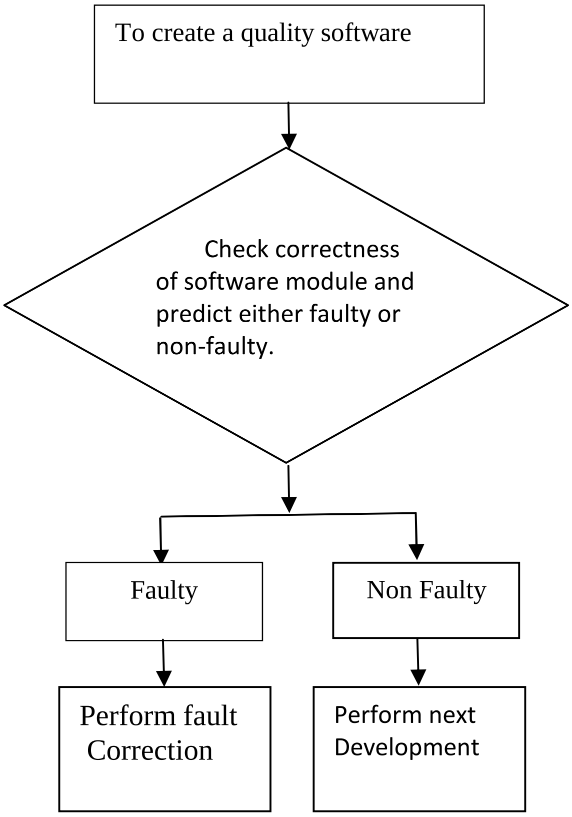
To create a quality software
Check correctness of software module and predict either faulty or non-faulty.
Faulty
Non Faulty
Perform fault Correction
Perform next Development

Fig.1: Software fault prediction model

The Figure 1 shows the software fault prediction model. It represents the correctness of software module and it verifies whether the software module is faulty. If it is faulty module then perform fault correction otherwise move to the next step. This paper concentrates only on fault prediction.

Data mining have two types of learning technique such as supervised and unsupervised learning technique. The class label of each training tuple is known is referred as supervised learning. Unsupervised learning represents the class label of each training tuple not known in advance [1]. Clustering is an unsupervised learning technique [12]. A cluster is a collection of data objects that are similar to one

Manuscript received Sep 15, 2012.

***Karpagavadivu.K**, Computer Science and Engineering, SNS college of Technology., (e-mail: kkarpagampg@gamil.com). Coimbatore, India.*

***Maragatham.T**, AP/Computer Science and Engineering, SNS college of Technology., (e-mail: tmaragatham@gmail.com). Coimbatore, India.*

***Dr.Karthik.S**, Dean/Computer Science and Engineering, SNS college of Technology., (e-mail: profskarthik@gmail.com). Coimbatore, India.*

another within the same cluster and are dissimilar to the objects in other clusters [1]. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering or cluster analysis [1]. Different clustering techniques are used to solve the problem. Clustering algorithms are divided into following categories:

1. Partitioning Methods
2. Hierarchical Methods
3. Density Based Methods
4. Grid Based Methods
5. Model Based Methods

II. Relevant Work

Open Source Software – Jedit:

Vikas Gupta et al. [5] surveyed that the basic concepts of clustering and they used metric values of JEdit open source software. In this paper, they formed a rules for categorization of software module is either faulty on non-faulty and empirically validation is performed. The results are measured in terms of accuracy of prediction, probability of detection and probability of false alarms. Finally, they conclude that open source software systems are analyzed. This model is implemented using K – Means based techniques for classification of software modules into faulty or non-faulty module [5].

X- Means Clustering Approach:

Catal. C et al. [3] proposed a new technique for software fault prediction. Their technique is only applicable to unlabeled program modules. They proposed a fully automatic technique and it is not needed to identify the number of clusters before clustering process starts like K -Means clustering method. This paper applied X – Means clustering method to cluster the modules and identifies the best cluster number. After this step, mean vector of each cluster is compared with metric threshold vector. If at least one metric value is higher than threshold metric value then that cluster is assigned as fault prone. They used three public data sets which locate in PROMISE repository [3].

Quad tree and EM Algorithm:

Meenu. S et al. [6] applied Expectation Maximization (EM) algorithm and Quad tree concept for predicting faulty modules. They found K – Means clustering algorithm has some drawbacks, so they propose one new algorithm (EM) and it is combined with Quad tree concept. Identify the centroid by Quad tree are input to EM algorithm. This algorithm gives the advantage of highest throughput than K – Means clustering algorithm, lesser number of iterations, lesser time and complexity. Finally they conclude that, it gives the clustering method not only fits the data better in clusters but also tries to make them compact and more meaningful. Their future work is to use HQ tree based EM clustering model. This HQ tree gives better cluster than Quad tree approach [6].

Decision tree and Fuzzy Logic:

AjeetKumar Pandey and Neeraj Kumar Goyal [7] surveyed that faults are predicted using data mining techniques and fuzzy logic. Decision tree is formed using ID3 algorithm (Iterative Dichotomiser). Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label [1]. The information gained from decision tree and that informations are converted into fuzzy rules. They propose Fuzzy Inference System (FIS) that system has input as software metrics and output is the degree of fault prone that decides whether module is fault prone or not. Aim of this paper is to help the software manager to improve the reliability and quality of software system [7].

RIDOR Algorithm:

Hassan Najadat et al. [13] proposed that a modification on RIDOR (RIpple DOwn Rule) algorithm that is they improved the effectiveness of RIDOR algorithm and that algorithm is refereed as Enhanced RIDOR algorithm. This enhanced algorithm learns defect prediction using mining static code attributes. These attributes are used to propose a new defect predictor with high accuracy and low error rate [13]. They use Weka tool for analyzing the data sets. The enhanced RIDOR has the benefit of two algorithms: CLIPPER and RIDOR. This paper used the rule based classification method for classification of modules from their fault prone. The goal of this paper is to improve the software development process and effectively allocate resources [13].

Metric Based Approach:

Shanthini.A and Chandrasekaran.RM [14] focused on high performance fault predictors that are based on machine learning algorithm. They used Method level metrics and Class level metrics for one type of data set. Support Vector Machine (SVM) provides the best prediction performance in terms of precision, recall and accuracy. Method level metrics are suitable for both procedural and object oriented programs. Class level metrics are only suitable for object oriented programs. They used four types of classifiers are: Naïve Bayes, K – Star, Random Forest and SVM. Their future work is to predict the software models based on some other machine learning algorithm [14].

Filter and Wrapper based Algorithms:

Akalya devi et al. [15] proposed a hybrid feature selection method which gives the better prediction than the traditional method. For evaluating the performance of software fault prediction models they used accuracy, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) [15]. In this paper they evaluated four filter based feature selection algorithms: 1. Correlation based feature selection(CFS) 2. Chi-squared 3. OneR 4. Gain Ratio. Also they evaluated three wrapper based feature selection algorithms: 1. Naive Bayes 2. RBF Network 3. J48 [15]. They conclude that, hybrid feature selection method gives better performance, reduce computational cost, reduce complexity of classifiers, better accuracy, improve productivity, easy maintenance and software quality.

III. CONCLUSION

The goal of this study is to analyze the performance of various techniques used in software fault prediction. And also, it describes some algorithms and its uses. Fault prone module prediction using data mining is to improve the quality of software development process. By using this technique, software manager effectively allocate resources. The overall error rates of all techniques are compared and we analyzed the advantages of all methods.

REFERENCES

[1]. Jiawei Han and Michline Kamber, " Data Mining concepts and techniques", Morgan Kaufmann publishers.

[2]. Maninderpalsingh, Nahasharma, Prabhjot Kaur, Varinderdeep Kaur, " Survey on software quality prediction models.

[3]. Catal. C, Sevim. U and Diri. B, " Software fault prediction of unlabeled program modules", WCE 2009, July 1-3, London,UK.

[4]. Sayward.F.G, Perlis.A.S and Shaw.M, " Software Metrics : Analysis and Evaluation", MIT press, Cambridge, 1981.

[5]. Parvinder S. Sandhu, Jagdeep Singh, Vikas Gupta, Mandeep Kaur, Sonia Manhas, Ramandeep Sidhu, **"**A K-Means Based Clustering Approach for Finding Faulty Modules in Open Source Software Systems", World Academy of Science, Engineering and Technology 72 2010.

[6]. Meenakshi P.C, Meenu S, Mithra M, Leela Rani P," Fault Prediction using Quad Tree and Expectation Maximization Algorithm", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Volume 2– No.4, May 2012 – www.ijais.org.

[7]. Ajeet Kumar Pandey, Neeraj Kumar Goyal," Predicting Fault-prone Software Module Using Data Mining Technique and Fuzzy Logic, Special Issue of IJCCT Vol. 2 Issue 2, 3, 4; 2010 for International Conference [ICCT-2010], 3rd-5th December 2010

[8]. Vijayalakshmi.M, Renuka Devi.M," A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets", Volume 2, Issue 3, March 2012 ISSN: 2277 128X.

[9]. Anoop Kumar Jain, Prof. Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining", Vol 1 Issue 1 Aug 2012 ISSN 2278-733X.

[10]. Pradeep Rai Shubha Singh, "A Survey of Clustering Techniques", International Journal of Computer Applications (0975 – 8887) Volume 7– No.12, October 2010.

[11]. Rajendra Prasad.K, Dr. Govinda Rajulu.P, "A Survey On Clustering Technique for Datasets Using Efficient Graph Structures", International Journal of Engineering Science and Technology Vol. 2 (7), 2010, 2707-2714.

[12]. Maryam hajiee, "A New Distributed Clustering Algorithm Based on K-means Algorithm", 2010 3rd International Conforence on Advanced Computer Theory and Engineering (1CACTE), pp. 408-411(Vol 2).

[13].Hassan Najadat and Izzat Alsmadi," Enhance Rule Based Detection for Software Fault Prone Modules", International Journal of Software Engineering and Its Applications Vol. 6, No. 1, January, 2012.

[14].Shanthini. A Chandrasekaran.RM,**"** Applying Machine Learning for Fault Prediction Using Software Metrics", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012, ISSN: 2277 128X.

[15]. Akalya devi.C, Kannammal. K.E and Surendiran.B," A Hybrid Feature Selection Model for Software Fault Prediction", International Journal on Computational Sciences & Applications (IJCSA) Vo2, No.2, April 2012.

[16] Arunachalam V.P, Karthik.S, "A Novel approach for mining inter-transaction itemsets", European Scientific Journal, 8(14).

Karpagavadivu.K received B.E degree in Computer Science and Engineering at Nandha Engineering College, Erode in 2005. She is currently pursuing M.E degree in Computer Science and Engineering at SNS College of Technology, Coimbatore. She has presented one international conference. Her areas of interests are Networks, data warehousing and data mining.

Maragatham. T received B.E degree in Computer Science and Engineering at M. Kumarasamy College of Engineering Karur in 2004. She has received Master Degree in Engineering from Anna University Coimbatore in 2011. She published papers in two international journals. She has presented papers in national and international conferences.

Professor Dr.S.Karthik is presently Professor & Dean in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University- Coimbatore, Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Anna University of Technology, Coimbatore. His research interests include network security, web services and wireless systems. In particular, he is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published more than 35 papers in refereed international journals and 25 papers in conferences and has been involved many international conferences as Technical Chair and tutorial presenter. He is an active member of IEEE, ISTE, IAENG, IACSIT and Indian Computer Society.