IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Software Fault Prediction using Data Mining Techniques: A Survey

Ashwni Kumar[1*], D.L. Gupta[2]

[1,2]Deptt of Computer Science and Engineering K.N.I.T, Sultanpur, U.P, India

[*]*Corresponding Author: Ashwni kumar_ashwanimpec20@gmail.com, Tel.: 8765001044*

DOI: https://doi.org/10.26438/ijcse/v7i6.671674 | Available online at: www.ijcseonline.org

Accepted: 13/Jun/2019, Published: 30/Jun/2019

***Abstract*—** In recent studies, it is found that a fault prediction technique plays an important role especially in software development. Software fault prediction implies a decent investment in better style in future system to avoid building a fault prone modules. Faulty modules are expected using data mining techniques such as various classifiers which are used to classify faulty or non faulty modules. Many researchers have been produces different approaches for predicting fault in the software. In this paper it is found that various fault prediction techniques have been used and also found out the way to judge the performance of fault prediction methodologies in recent year. The main objective of survey is to identify best prediction techniques for detecting fault in early stage, and also determine the problem area in software fault prediction methodology which provides improvement in software development system. This paper presents the survey on fault prediction using data mining techniques which will helpful for further research in field of software fault prediction.

***Keywords*—** Software fault prediction, Data Mining, Prediction techniques.

I. INTRODUCTION

A software defect is a fault, error, or failure in software [1].It produces either unexpected or incorrect result or it act like unintended ways. It is an inadequacy in software product that causes it to perform unusually. As the present software's are develop quickly because of the size and complexity, in software development process software reviews and testing is the most important phase of finding software defect.

As software is developed by human being, it may be have a lot of defect can be generated during software development life cycle. It is very hard to develop fault free software, quality software. It is, therefore of great worry to locate fault prone software modules at an early stage of the software project. Finding the fault as soon as possible in software development process will not only improve the effective cost but it also helps to achieve customer satisfaction. It is very necessary to predict the faults in the software because it helps in reducing the cost and development of good quality of reliable software. The defective modules in the software can be predicted by software fault prediction techniques. So to find out the faults we need test data which can measure the faults from system at various phases of software life cycle.

II. LITERATURE SURVEY

In software development life cycle, software is existing in different forms. So it may be software Requirements, specification document in the analysis phase, or a design pattern in the design phase, or executable software. A various number of software fault Prediction models and techniques have been proposed by different researchers in last few year.

Two-Stage Data Preprocessing Approach:
Wangshu Liu et al. [2] has been proposed for a two-stage data processing technique, which provides better instance reduction process and feature selection [2]. By using classification model for software fault prediction the quality of software dataset have been improved for better prediction. In the feature selection process, we suggest a novel algorithm, which provides both redundancy control and relevance analysis. And in the instance reduction process, we recommend random under sampling to keep the faulty and non faulty instance balanced.

A Hybrid Approach:
Yasutaka Kamei et al. [3] have been proposed for a method for Hybrid faulty prone module prediction to combine logistic regression and association rule mining analysis together. If a module satisfies the conditions on basis of one of the selected rules, the module is classified by rule as either

faulty or not. Otherwise, if condition is not satisfied the module the logistic regression classified is applied. The performance of prediction of this model was evaluated and compared with three other fault prone modules based on logistic regression model, linear discriminate model and classification tree. The results of experiments showed improvement in performance as compared to conventional methods.

Clustering Approach:
Catal. C et al. [4] proposed an X- Means Clustering technique for software fault prediction. Their procedure is just appropriate to unlabeled program module. They proposed a completely automatic technique and it is not required to recognize the number of clusters before clustering process begins like K - Means clustering method. In this paper it applied X-Means clustering technique to the cluster module and distinguishes the best cluster number. After previous process the next step is that, mean vector of each cluster is compare with metric threshold vector. The cluster is assigned fault prone only if it satisfied the condition of at least one metric value is higher than threshold metric value. They used three public dataset which is present in PROMISE repository [4].

Metric Based Approach:
Shanthini.A et al. [5] proposed an approach based on machine learning algorithm for high performance fault prediction. They used Method level metrics and Class level metrics for one type of data set. Support Vector Machine provides the best prediction performance in terms of precision, recall and accuracy. Method level metrics are applicable for both object oriented programs and procedural programs. Class level metrics are only suitable for object oriented programs. They used four types of classifiers are: Naive Bayes, K – Star, Random Forest and SVM. Their future work is to predict the software models based on some other machine learning algorithm [5].

Quad tree and EM Algorithm:
Meenu. S et al. [6] has proposed a technique based on Quad tree algorithm and Expectation Maximization algorithm for predicting fault modules in the software. They found K–Means clustering algorithm has some disadvantages, so to remove these disadvantages they proposed a new concept of combining of EM algorithm and Quad tree algorithm. Identify the centroid by Quad tree are input to EM algorithm. This algorithm gives the advantage of highest throughput as compare to K – Means clustering algorithm, and it has less number of iterations, and it's time complexity is also less then K-Mean clustering algorithm [6].

Class Association Rule Approach:
Yuanxun SHAO et al. [7] have proposed a new approach for software defect prediction based on using Class Association Rule. In this approach class association rule is treated as a independent class label, which is a particular type of association rule that define the relation between attributes and categories for defect perdition.

III. DATA MINING TECHNIQUES FOR FAULT PREDICTION

Data mining is of great importance since 1990s. The main objective of the data mining method is extracting and analysis of valuable information from massive data is the key objective of Data Mining Technique [8].There are various data mining techniques are used for fault prediction and still lot of research is going on to find out new techniques that can produce better outcomes. There are some following Data Mining Techniques which are explained in brief below [8] [9].

Classification
Classification is that the process of generalization used to classify every item in a set of data into one in all predefines set of cluster or classes. An algorithm which implements the classification is known as classifier. Classifier term is referring as mathematical function. The main operation of classification technique could be data processing operation that assigns items in a collection to focus on classes. The objective of classification technique is to predict the target class of every case within the data [10].

Decision Tree
The Decision Tree is classification approach in which classification is done diving rule. The decision tree is formation like a tree structure that groups instances by sorting them in aspect of the feature values. Decision tree makes the rule for the classification of data set. In decision tree every node in the tree represents a feature and the branch represents the value. Classification starts at the root node and moves to the leaf node for prediction of the class that a particular instance belongs to. The three fundamental algorithms are broadly utilized that are ID3, CART, and C4.5.

a. ID3:- ID3 stands for (Iterative Dichotomiser 3**)** is an algorithm invented by Ross Quinlan .It is a classification algorithm. Its essential thought is that all models are summarized to distinctive categories according to distinctive values of condition attribute set; its base is to decide on the simplest classification attribute from condition attribute sets.

b. CART:-CART stands for classification and regression tree. This algorithm was invented by Breiman in 1984.The development of classification tree is totally depend on binary separating of the properties. CART approach is based on Hunt's algorithm and can be

executed serially. CART uses data gain as splitting criterion. The topmost decision node is that the simplest and the most effective predictor, it's referred to as root node. The attribute with highest data Gain is selected as split attribute. Data gain is employed to form tree from training instances. This tree is utilized as classify test data [11].

c. **C4.5:-** C4.5 algorithm was given by Ross Quinlan in 1993, it is a modified version of ID3 induction algorithm. The decision tree generated by C4.5 algorithm is utilized for classification is usually referred to as a statistical classifier.C4.5 algorithm accepts data with numerical values or categorical values. The C4.5 is mostly use as free data mining tool.

Naïve Bayesian (naïve Bayes)

Naïve Bayes is simplest classifier which is relied on Bayesian theorem [12]. The Bayesian classification considered as supervised learning strategy and statistical techniques for classification. This classifier is utilized for both the feature which is independent to every class and also for those features where independence is no further valid. Naïve Bayes technique work in two stages: training stage and prediction stage.

Support vector machine (SVM)

SVM were first suggested by Vapnik in the 1960.Support Vector Machines are supervised learning model and when it combines with associated learning algorithm then it provide the complete analysis of data and recognize information patterns. For supervised learning model of classification and regression rules Support Vector Machine is used as training algorithm for the data. To solve the classification problem of supervised learning it required a tool which is known as support vector machine.

Regression

Regression is method by which we can easily identify those function that are helpful in order to determine the correlation among different various variables. It is basically used to find out the function that explains the correlation between different variables. It is basically mathematical tool, used for constructing training dataset. Regression is statistical method which examines the relationship between variable. In statistical method two types of variables are used where one is called dependent variable and another variable is independent variables and it is represented by 'Y' and 'X'. In regression there is always one dependent variable while independent variable may be one or more than one.

Clustering

Clustering is the method of grouping of an object into similar type of class objects. Clustering is the technique by which the grouping of data into clusters, so those objects with in a cluster have much similarity in comparison to each other but it is different from objects in distinct clusters. For distinguishing dense and distributed region in an object space clustering technique is employed and that we discover distribution pattern and confirm the correlation among data attributes. In general, the clustering techniques can be classified into the subsequent categories.

a. Partitioning method
b. Hierarchical method
c. Density based method
d. Model based method
e. Grid based method

Association Rule Mining

In data mining the most important technique is used as Association Rule mining. It is the method of finding frequent large data set with the help of association and correlation. Association Rule mining can predict the simplest occurrence of item sets supported the occurrence of different items sets.

IV. CONCLUSION AND FUTURE SCOPE

After study of various researches related to data mining technologies like classification, regression, clustering, and association rule have been used for software fault prediction. We made a theoretical and methodological observation during study of these methods for fault prediction, and after that we have found following shortcoming which was concluded in these papers. In previously reviewed paper they have been completely utilizing PROMISE repository data set, as a new contains pre processed data set so, therefore which has to be changed in pre processing techniques.In the most of previous researchers they have only focused on using single data mining techniques like classification, regression and etc so, we request for further research to use combined (Hybrid) prediction techniques, by which they can obtain best possible results in software fault prediction.

REFERENCES

[1] Naik, K., & Tripathy, "*Software Testing and Quality Assurance*", John Wiley & Sons, Inc, pp. (2008).

[2] 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.

[3] 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**.

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

[5] 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.**

[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

[7] Yuanxun SHAO, Bin LIU, Guoqi LI, Shihai WANG," *Software Defect Prediction Based on Class Association Rules*", The Second International Conference on Reliability Systems Engineering (ICRSE **2017**)

[8] N.P. Gopalan, B. Sivaselvan, *"Data Mining Techniques and Trends"*, PHP Learning Private Limited, New Delhi-110001, **India**, ISBN-**978-81-203-3812**-B, **2009**

[9] G.K. Gupta, "*Introduction To Data Mining With Case Studies*", Second Edition, PHP Learning Private Limited, New Delhi-110001, **India**, ISBN-**978-81-203-4326-9, 2011**

[10] D. Hand, H. Mannila and P. Smyth, "*Principles of data mining*", MIT, **2001**.

[11] Fong, P.K. and Weber-Jhanke, J.H , *"Privacy Preserving Decision Tree Learning using Unrealized Data Sets"*, IEEE Transactions on knowledge and Data Engineering, Vol.**24**,No.**2**, February **2012**, pp. **353-364**.

[12] Langley P, Iba W, Thompson K., "*An analysis of Bayesian classifiers",* in Proceedings of the 10th National Conference on Artificial Intelligence, **1992**, pp. **223-228**.

