A Study on Software Defect Prediction by Data Mining Techniques

Sonu Kumar Kushwaha
M.Tech Scholar, Department of CSE
SIRT, Bhopal INDIA
E-mail- sonukushwaha29@gmail.com

Prof. Sunil Malviya
Department of CSE
SIRT, Bhopal INDIA
E-mail- sm.sunil84@gmail.com

ABSTRACT
Software defect prediction work focuses on the number of defects remaining in a software system. The software defect prediction model helps in early detection of defects and contributes to their efficient removal and producing a quality software system based on several metrics. A prediction of the number of remaining defects in an inspected are fact can be used for decision making. An accurate prediction of the number of defects in a software product during system testing contributes not only to the management of the system testing process but also to the estimation of the product's required maintenance. Defective software modules cause software failures, increase development and maintenance costs, and decrease customer satisfaction. It strives to improve software quality and testing efficiency by constructing predictive models from code attributes to enable a timely identification of fault-prone modules. The main objective of paper is to help developers identify defects based on existing software metrics using data mining techniques and thereby improve the software quality. In this paper, we will discuss data mining techniques that are association mining, classification and clustering for software defect prediction. This helps the developers to detect software defects and correct them.

Keyword: - Software defect prediction, data mining, clustering, classification and association rule mining.

INTRODUCTION
In context of software engineering, software quality refers to software functional quality and software structural quality. Software functional quality reflects functional requirements whereas structural quality highlights non-functional requirements. Software quality measurement [1] is about quantifying to what extent a system or software possesses desirable characteristics namely Reliability, Efficiency, Security, Maintainability and (adequate) Size. This can be performed through qualitative or quantitative means or a mix of both. In both cases, for each desirable characteristic, there are a set of measurable attributes like Application Architecture Standards, Coding Practices, Complexity, Documentation, Portability and Technical & Functional Volumes. The existence of these attributes in a piece of software or system tends to be correlated and associated with this characteristic. A software defect is an error, flaw, mistake, failure, or fault in a computer program or system that produces incorrect or unexpected results, or causes it to behave in unintended way. Software defect prediction is the process of locating defective modules in software. It helps to improve software quality and testing efficiency by constructing predictive models from code attributes to enable a timely identification of fault-prone modules, it also helps us in planning, monitoring and control and predict defect density and to better understand and control the software quality. The Software Defect Prediction result, that is the number of defects remaining in a software system, it can be used as an important measure for the software developer, and can be used to control the software process [2].

In this paper, we will discuss Data mining techniques for software defect prediction. Data mining is a process of analysing data from different perspectives and summarizing it into useful information. It allows users to understand the substance of relationships between data. It reveals patterns and trends that are hidden among the data. It is viewed as a process of extracting valid, previously unknown, non-trivial and useful information from large databases. The techniques of data mining for software defect prediction are: clustering, association mining, and classification. In rest of the paper Section 2 presents the related work on the topic, Section 3 presents the data mining techniques for Defect Prediction model and at last Section 4 presents the conclusion and future work.

II BACKGROUND OF WORK
In [3] Q. Song, et. al. presented an application of association rule mining to predict software defect associations and defect correction effort with SEL defect data. This is important in order to help developers detect software defects and project managers improve software control and allocate their testing resources effectively. The objective of the study is to discover software defect associations from historical software

engineering data sets, and help determine whether or not a defect is accompanied by other defects. If so, we attempt to determine what these defects are and how much effort might be expected to be used when we correct them.

In [4], they applied K-Means and Neural-Gas techniques on different real data sets and then the representative module of the cluster and several statistical data are explored in order to label each cluster as fault-prone or not fault-prone. In their study they have presented comparative results performed on same data sets. They have applied unsupervised learning approach for fault prediction in software module. The false negative rates (FNR) for the clustering-based approach are less than that for metrics-based approach, while the false positive rates (FPR) are better for the metrics-based approach. The overall error rates for both approaches remain the same. Quad Trees are applied for finding the initial cluster centres for K-Means algorithm. The overall error rates of the software fault prediction approach by Quad Tree based K-mean algorithm are found comparable to other existing algorithms.

In [5], predictive models are estimated based on various code attributes to assess the likelihood of software modules containing errors. Many classification methods have been suggested to accomplish this task. In this paper, they assess the use of classification method, CBA2, and compare it to other rule based classification methods. They have investigated the performance of an association rule based classification method for software defect prediction problems. Data experiments were conducted to compare the CBA2 classifier with two other rule/tree based classifiers showing that the CBA2 method obtained satisfactory performance when compared to C4.5 and RIPPER.

III SOFTWARE DEFECT PREDICTION BY DATA MINING TECHNIQUES

Data mining is the analysis step of the "Knowledge Discovery in Databases" process, or KDD, a process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems[6]. Data Mining can be divided into two tasks: Predictive tasks and descriptive tasks. Predictive task is to predict the value of a specific attribute (target/dependent variable) based on the value of other attributes (explanatory). Descriptive task is to derive patterns (correlation, trends, and trajectories) that summarize the underlying relationship between data. In this paper, different data mining techniques are discussed for identifying fault prone modules. Data Mining plays an important role in software defect prediction. It helps in cleaning data. For this the data is taken from the Software repositories. It has lots of information that is useful in assessing software quality. Data mining techniques and can be applied on these repositories to extract the useful information. Data mining techniques can be applied on the software repositories to extract the defects of a software product.

1. Regression:

It is a statistical process to evaluate the relationship among variables. It analyses the relationship between the dependent or response variable and independent or predictor variables. The relationship is expressed in the form of an equation that predicts the response variable as a linear function of predictor variable [7-10].

Linear Regression: $Y=a+bX+u$.

2. Clustering:

Clustering is a form of unsupervised learning in which no class labels are provided. It is often the first data mining task applied on a given collection of data. In this, data records need to be grouped based on how similar they are to other records. It is a task of organizing data into groups such that the data objects that are similar to each other are put into same cluster. The groups are not predefined. It is a process of partitioning a data in a set of meaningful sub-classes called clusters. Clusters are subsets of objects that are similar. Clustering helps users to understand the natural grouping or structure in a data set. Its schemes are evaluated based on the similarity of objects within each clusters. In [11, 4], k-mean technique of clustering used for Software Defect Prediction. K-mean clustering is a non-hierarchical clustering procedure in which items are moved among sets of clusters until the desired set is reached. It has certain drawbacks, so to overcome those drawbacks Quad Tree-based k-mean clustering method was proposed. The objective was: first, Quad-trees are applied for finding initial cluster centres for k-mean algorithm. Second, the Quad tree-based algorithm is applied for predicting faults in program modules. They have evaluated the effectiveness of Quad tree-based k-mean clustering algorithm in predicting faulty software modules as compared to the original k-mean algorithm. The result of this Quad tree-based k-mean algorithm is compared with other approaches and found that the number of iterations of k-means algorithm is less in case of Quad tree-based k-mean except for other approaches, as well as percent error also give fairly acceptable values.

3. Classification:

Classification is a process of finding a set of models that describe and distinguish data classes or concepts. It is the organization of data in given classes known as supervised learning, where the class labels of some training samples are given. These samples are used to supervise the learning of a classification model. In classification test data are used to estimate the accuracy of the classification rules. A large number of classification methods have been suggested to build software defect prediction models. In [5], an association rule classification method is proposed which derives a comprehensible rule set from the data. They have compared CBA2 [12] with other rule based classification method to check whether classification algorithms based on association rules are suitable for software fault prediction or not. They have also tried to find out whether rule sets learned on one data set are applicable to others data sets or not. They have investigated the performance of an association rule based classification method for software defect prediction. The experiments were conducted on the data sets and the result

was compared with other classifiers and obtained the satisfactory performance without losing comprehensibility.

4. Association Mining:

The Association mining task consists of identifying the frequent item sets, and then forming conditional implication rules among them. It is the task of finding correlations between items in data sets. Association Rule algorithms need to be able to generate rules with confidence values less than one. Association rule mining is undirected or unsupervised data mining over variable-length data and it produces clear, understandable results. The task of association rules mining consists of two steps. The first involves finding the set of all frequent itemsets. The second step involves testing and generating all high confidence rules among itemsets. In association rule mining technique we use defect type data to predict software defect associations that are the relations among different defect types. The defect associations can be used for three purposes: First, Find as many related defects as possible to the detected defects and make more effective corrections to the software. Second, it helps to evaluate reviewer's results during an inspection. Third, it helps in assisting managers in improving the software process through analysis of the reasons some defects frequently occur together. Association rule mining aims at discovering the patterns of co-occurrences of the attributes in the database. They found that higher support and higher confidence levels may not result in higher prediction accuracy.

IV CONCLUSION

In this paper, we have discussed that how data mining techniques are used for software defect prediction. In order to improve the efficiency and quality of software development, we can make use of the advantage of data mining to analysis and predict large number of defect data collected in the software development. We have also studied in previous papers that how these techniques have performed better results when performed on different data sets. In future, we will be be comparing the results of the techniques discussed above to see which is better.

REFERENCES:-

[1]. The Global Conference for Wikimedia,(2014); London.

[2]. P.J. Kaur ,Pallavi , " Data Mining Techniques for Software Defect Prediction ", International Journal of Software and Web Sciences 3(1), December, 2012-February, 2013, pp. 54-57.

[3]. Qinbao Song, Martin Shepperd, Michelle Cartwright, and Carolyn Mair, "Software Defect Association Mining and Defect Correction Effort Prediction", IEEE Transactions on software engineering, Vol. 32, no. 2, February 2006.

[4]. Partha Sarathi Bishnu and Vandana Bhattacherjee, "Software Fault Prediction Using Quad Tree-Based K-Means Clustering Algorithm", IEEE Transactions on knowledge and data engineering, Vol. 24, no. 6, June 2012.

[5]. Baojun Ma1 Karel Dejaeger2 Jan Vanthienen2 Bart Baesens2, "Software Defect Prediction Based on Association Rule Classification", The 2010 International Conference on E-Business Intelligence.

[6]. M.C.M. Prasad, L.Florence,A.Arya," A Study on Software Metrics based Software Defect Prediction using Data Mining and Machine Learning Techniques" International Journal of Database Theory and Application Vol.8, No.3 (2015).

[7]. R.Goyala, P.Chandraa, Y. Singha, "Suitability of KNN Regression in the Development of Interaction Based Software Fault Prediction Models", IERI Procedia, International Conference on Future Software Engineering and Multimedia Engineering, Elsiever, vol 6, pp. 15-21, (2013).

[8]. G.Scanniello, C.Gravino, A.Marcus,T.Menzies,"Class level fault prediction using software clustering, Automated Software Engineering (ASE)", 2013 IEEE/ACM 28th International Conference, (2013).

[9]. M. Jureczko, "Significance of Different Software Metrics in Defect Prediction", Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology,WybrzeżeWyspiańskiego vol. 27, pp.50-370.

[10]. S.Bibi, G.Tsoumakas, I.Stamelos, I. Vlahavas, "Software Defect Prediction Using Regression via Classification", Department of Informatics, Aristotle University of Thessaloniki,54124 Thessaloniki, Greece.

[11]. Michael Laszlo and Sumitra Mukherjee, Member, IEEE, "A Genetic Algorithm Using Hyper-Quadtrees for Low-Dimensional K-means Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, no. 4, April 2006.

[12]. B. Liu, Y. Ma, C.K. Wong, "Improving an association rule based classifier," In Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, 2000.