



## Data Mining Techniques for Software Defect Prediction

Ms. Puneet Jai Kaur<sup>1</sup>, Ms. Pallavi<sup>2</sup>

<sup>1</sup>Assistant Professor, UIET, Panjab University, Chandigarh, India

<sup>2</sup>Student, UIET, Panjab University, Chandigarh, India

**Abstract-** Software defect prediction work focuses on the number of defects remaining in a software system. A prediction of the number of remaining defects in an inspected artefact 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. 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. Unsupervised techniques may be used for defect prediction in software modules, more so in those cases where defect labels are not available.

**KEYWORDS:** software defect prediction; data mining; clustering; classification; and association rule mining.

### I. Introduction

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. 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 association mining, we use rules to predict software defect associations that are the relations among different defect types such as: If defects a and b occurs, then defect c also will occur. This is formally written as  $a \wedge b \Rightarrow c$ . Association Rule mining is not restricted to dependency analysis in the context of retail applications, but can successfully be applied to a wide range of business and science problems. Although association rule mining aims to discover the co-occurring patterns of the attributes in databases, it has been shown that association rule mining based classification—associative classification—frequently has higher classification accuracy than other classification methods.

Clustering is a non-hierarchical procedure in which items are moved among sets of clusters until the desired set is reached. Each clustering technique makes some assumptions about the underlying data set. If the assumptions hold, good clustering can be expected. But, it is hard to satisfy all the assumptions. Therefore, it is beneficial to apply different clustering methods on the same data set, or the same method with varying input parameters or both.

Classification is a process of finding a set of models that describe and distinguish data classes or concepts. The derived model is represented in various forms such as classification rules, decision tree etc. Once classes are defined the system should infer rules that govern the classification. Therefore the system should be able to find the description of the each class. The description should only refer to the prediction attributes of the training set so that the positive examples should satisfy the description. A rule is said to be correct if its description covers all the positive examples and none of the negative examples of a class.

The rest of the paper is organized as follows: Section 2 presents the related work on the topic. Section 3 presents the data mining techniques for Defect Prediction model. Section 4 presents the conclusion and future work.

## II. Related Work

In [1], 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 [2] we have 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. We aim to help detect software defects and effectively improve software control.

In [3], 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. Data Mining Techniques for Defect Prediction Models

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 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.

### A. 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 organising 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. This approach offers benefits to the expert who must decide the labels. Instead of inspecting and labelling software modules one at a time, the expert can inspect and label a given cluster as a whole; he or she can assign all the modules in the cluster the same quality label.

- **Software Defect Prediction using Clustering**

In [1], [4], they have used k-mean technique of clustering 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.

## **B. 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. Classification approaches normally use a *training set* where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules.

- **Software Defect prediction using classification**

A large number of classification methods have been suggested to build software defect prediction models. In [3], an association rule classification method is proposed which derives a comprehensible rule set from the data. They have compared CBA2 [9] 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.

## **C. Association Mining**

The Association mining task consists of identifying the frequent itemsets, 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. It has a simple problem statement, that is, to find the set of all subsets of items that occur frequently in many database records or transactions, and to extract the rules telling us how a subset of items influences the presence of another subset.

- **Software Defect prediction using association mining**

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 and future work**

In this paper, we have discussed that how data mining techniques are used for software defect prediction. 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 comparing the results of the techniques discussed above to see which is better.

## **V. References**

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

- [2] 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.
- [3] Baojun Ma<sup>1</sup> Karel Dejaeger<sup>2</sup> Jan Vanthienen<sup>2</sup> Bart Baesens<sup>2</sup>, "Software Defect Prediction Based on Association Rule Classification", The 2010 International Conference on E-Business Intelligence.
- [4] Michael Laszlo and Sumitra Mukherjee, Member, IEEE, "A Genetic Algorithm Using Hyper-Quadrees for Low-Dimensional K-means Clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, no. 4, April 2006.
- [5] Lourdes Pelayo and Scott Dick, "Evaluating Stratification Alternatives to Improve Software Defect Prediction", *IEEE Transactions on Reliability*, Vol. 61, no. 2, June 2012.
- [6] Qinbao Song, Zihan Jia, Martin Shepperd, Shi Ying, and Jin Liu, "A General Software Defect-Proneness Prediction Framework", *IEEE Transactions on Software Engineering*, Vol. 37, no. 3, May/June 2011.
- [7] Naheed Azeem, Shazia Usmani, "Analysis of Data Mining Based Software Defect Prediction Techniques", *Global Journal of Computer Science and Technology*, Volume 11 Issue 16 Version 1.0 September 2011.
- [8] [en.wikipedia.org/wiki/software\\_bug](http://en.wikipedia.org/wiki/software_bug).
- [9] 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.
- [10] Stefan Lessmann, Student Member, IEEE, Bart Baesens, Christophe Mues, and Swantje Pietsch, "Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings", *IEEE Transactions on Software engineering*, Vol. 34, no. 4, JULY/AUGUST 2008.