**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**

*IJCSMC, Vol. 3, Issue. 5, May 2014, pg.879 – 886*

**REVIEW ARTICLE**

# A Review on Software Defect Prediction Models Based on Different Data Mining Techniques

**Raminder Kaur[1], Punam Bajaj[2]**

[1], [2] Department of CSE, Chandigarh Engineering College, Mohali, Punjab, India

[1] raminder.kaur1077@gmail.com; cecm.cse.punb@gmail.com [2]

***Abstract—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. Empirical studies have been carried out and a number of approaches have been developed and proposed over the years, and many models have been proposed with different predictive capabilities and efficiency. This paper presents the survey on existing data mining techniques used for prediction of software defects. This paper will also introduce the concept of neural networks which is been considered as one of the promising technique for predictive models.***

***Keywords: Data Mining; Software Defect Prediction; Software Reliability***

## 1. INTRODUCTION

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. 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. 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.

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.

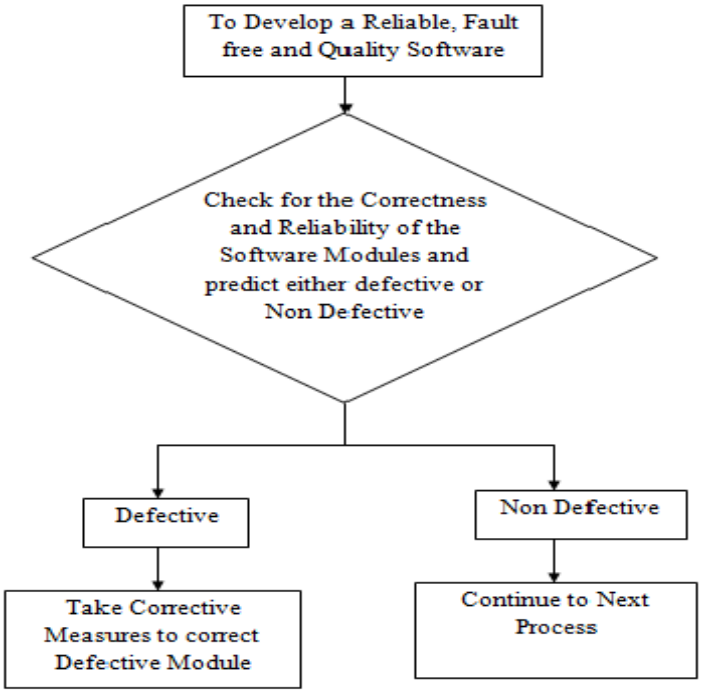


Figure 1: Software Defect Prediction

## 2. DATA MINING & ITS TECHNIQUES USED FOR PREDICITION

Data mining is a process to convert raw data into useful information. It is a process designed to explore and analyse large amounts of data to find consistent patterns, trends or relationships among variables and then to validate the findings by applying the observed patterns to new set of data.

Data Mining can be divided into two tasks: Predictive tasks and descriptive tasks. The ultimate aim of data mining is prediction and predictive data mining is the most common type of data mining and has the most direct applications to business [28]. Predictive Data mining relies on formulas that compare pass successes and failures, and then uses those formulas to predict future outcomes.

Data Mining is about explaining the past and predicting the future by means of data analysis. It is multidisciplinary field which combines statistics, machine learning, artificial intelligence and database technology [29]. Data mining techniques aids predictive analytics to analyse historical facts in order to predict about future events. Predictive Modelling is the process by which a model is created to predict an outcome. If the outcome is categorical it is called classification and if the outcome is numerical it is called regression. Descriptive modelling or clustering is the assignment of observations into clusters so that observations in the same cluster are similar. Finally, association rules can find interesting associations and correlations amongst observations [29].

 

There are various data mining techniques used for predictions which are discussed below.

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.
   Linear Regression: Y=a+bX+u

2. Association Rule Mining: It is a method for discovering interesting relationships between variables in large databases. It is about finding association or correlations among sets of items or objects in database. It basically deals with finding rules that will predict the occurrence of item based on the occurrence of other items.

3. Clustering: Clustering is a way to categorize a collection of items into groups or clusters whose members are similar in some way. It is task of grouping a set of items in such a way that items in the same cluster are similar to each other and dissimilar to those in other clusters.

4. Classification: It consists of predicting a certain outcome based on a given input. Classification technique use input data, also called training set where all objects are already tagged with known class labels. The objective of classification algorithm is to analyze and learns from the training data set and develop a model. This model is then used to classify test data for which the class labels are not known.
   a. Neural Networks: Neural Networks are the non linear predictive models which can learn through training and resemble biological neural networks in structure. A neural network consists of interconnected processing elements called neurons that work together in parallel within a network to produce output.

   b. Decision Trees: A decision tree is a predictive model which can be used to represent both classification and regression models in the form a tree structure. It refers to a hierarchical model of decisions and their consequences. It is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf nodes represent a classification or decision.

   c. Naive Bayes: It is based on Bayes theorem with independence assumption between predictors. Naive Bayes Classifier is based on the assumption that the presence or absence of a particular feature of a class in not related to the presence or absence of any other feature.

   d. Support Vector Machines: SVM are based on the concept of decision planes that define decision boundaries. A decision plane is the one that separates between a set of objects having different class membership. SVM is primarily a classifier method that performs classification task by constructing hyper plane in a multidimensional space that separates cases of different class labels. It supports both regression and classification.

   e. Case Based Reasoning: Case based reasoning means solving new problems based on the similar past problems and using old cases to explain new situations. It works by comparing new unclassified records with known examples and patterns. A simple example of a case based learning algorithm is k-nearest neighbour algorithm. It is simple algorithm that stores all available cases and classifies new cases based on a similarity measure i.e. distance function.

## 3. LITERATURE SURVEY

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.

 

### *A. Software Defect Prediction Based on Classification Techniques*

Karunanithi et.al[11] 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.

Khoshgaftaar et.al[21] introduced the use of the neural network as a tool for predicting software quality of a very large telecommunication system, classifying modules into fault or non fault prone. They compared the Artificial Neural Network model with a non parametric discriminant model, and found that Neural Network model has better predictive accuracy.

Kanmani et al[19] introduced two neural network based fault prediction models using object oriented metrics and compared the results with two statistical models using five quality attributes and concluded that neural networks do better.
In [15], researchers have done a comparison of clustering based approach and neural based approach on real time data set and found that performance is better in case of neural network approach in terms of accuracy, mean absolute error and root mean square error values.

Some researchers also proposed a non parametric software reliability prediction systems based on neural network ensembles. In [4] and [12], researchers combined a number of different neural networks and presented the significant improvement in performance of software reliability forecasting over individual neural network based model.

Khoshgaftaar et.al [22] presented a software fault prediction approach with case based reasoning. In addition it was also concluded that CBR models have better performance than models based on multiple linear regression.

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 and See5.

In [9], an association based classification method; CBA2 is proposed and is compared to other classification methods. The Results found showed that applying the CBA2 algorithm result in accurate rule sets. Data experiments were conducted to compare the CBA2 classifier with two other rule based classifiers showing that the CBA2 method obtained satisfactory performance when compared to C4.5 and RIPPER

Authors in [13] have compared the performance of prediction models by using static attributes of embedded software. Three machine learning algorithms i.e. J48, OneR and Naïve Bayes have been used for prediction purpose of two datasets. It was found that J48 and OneR performed better than Naïve Bayes learner.

Authors in [5] have applied Support Vector Machines for predicting fault prone software modules and its prediction performance is compared against eight statistical and machine learning methods in the context of four NASA datasets. The results indicate that prediction performance of SVM is generally better than the compared models

In [18], defect prediction is performed using method level metrics and class level metrics for KC1 dataset. Various classifiers like Naive Bayes, Support Vector machines, K-Star, Random forest were examined using traditional measures such as accuracy, precision, recall and F-measure. The SVM Method outperforms other classifiers for class level metrics and Random forest shows better performance for method level metrics

 

In [20], researchers proved that classifier ensembles can effectively improve classification performance than single classifier. A Comparative study of various ensemble methods was conducted. These methods include Bagging, Boosting, Random Trees, Random Forest, Random Subspace, Stacking and Voting. These ensemble methods were also compared to a single classifier Naïve Bayes and showed that applying ensemble methods could achieve better performance than using a single classifier.

A detailed analysis of some of the existing machine learning techniques for defect prediction has been carried out using four different data sets from NASA MDP repository [23]. The techniques considered for experiment included Decision tree, Naive Bayes Class Classifier, Logistic Regression, Support Vector Logistic regression, Neural Network for discrete goal field(NND), 1-Rule, Instance Based Learning for 10 nearest neighbours(IBL). Mean absolute error was considered for performance assessment. Based on analysis, Naive Bayes Classifier, IBL, NND performed better than other prediction models. Moreover the paper also concluded that the selection of best learning techniques depends on data insight at that point in time.

In [1], researchers have used software size metrics and three metrics of requirement analysis phase for predicting the software defect using fuzzy logic and examined the predictive quality of the proposed approach using qualitative data of requirement metrics of twenty real software projects.

***B. Software Defect Prediction Based on Clustering Techniques***

In the paper [24], 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%.

In the paper [3], 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% 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.

***C. Software Defect Prediction Based on Association rule mining***

In [14], 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.

In [2], 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.

***D. Software Defect Prediction Based on Hybrid Approach***

In the paper[7], 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.

Hybrid fault prone module prediction method was introduced that combines association rule mining with logistic regression analysis [26]. 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

## 4. DISCUSSION

On reviewing literature, it is found that various machine learning approaches such as supervised and unsupervised learning have been used for building a software defect prediction models. Among these, supervised learning approach is extensively used and found to be more useful for predicting software defects if sufficient amount of previous fault data is available. The Supervised learning approach cannot be used effectively to build powerful models when previous data available is limited. An effective and low cost method to predict defects is learning from previous mistakes and prevent in future.

Machine learning has emerged as a significant way to predict the existence of defects in the software modules. There are two essential steps involved in machine learning approach. First, learning knowledge from the previous existing data; secondly, predicting the future (on new data).The Classifier is first trained using software fault history data and then used to predict faults. Major Algorithms used in machine learning are decision trees, artificial neural networks ,Regression analysis, case based reasoning, Support Vector Machines, Random forest, discriminant analysis, fuzzy classification, Bayesian Belief networks, bagging and boosting.

Existing research in software defect prediction models focus on predicting faults from these perspectives:

- The Number of Faults: This approach predicts the number of faults in a software module.
- Classification: Classification models are used to predict defects in a software module. They basically classify the modules into faulty and non faulty. This kind of prediction distinguishes fault free systems from fault prone systems.

### *A. Role of Artificial Neural Networks in Software Defect Prediction*

Software defect prediction models are mostly based on relationship between the software metrics and fault proneness. However, this relationship is often complex and non linear. This limits the usefulness of conventional statistical approaches.
Artificial Neural Networks have gain immense popularity over the years because it is very sophisticated modelling technique and can model very complex functions and non linear relationships that are difficult to model with other techniques and thus are applicable for software quality modelling.

Neural Networks models have significant advantage over analytical models because they require only failure history as input, no assumptions. Using that input, neural network model automatically develops its own internal model of failure process and predicts future failure. Neural Network is a collection of fast processing and interconnected computing nodes called neurons. Each neuron can receive signal, process the signals and finally produce an output signal.

TABLE 1: Pros and Cons of Neural Networks

| Sr. No. | Pros | Cons |
|---|---|---|
| **1.** | Neural Networks are the closest thing to have an actual human operating a system i.e. they can learn. | Neural Networks are difficult to design as we have to determine the optimal number of nodes, hidden layers, activation function etc**.** |
| **2.** | Learning from the past history data is the most important approach of ANN which acts as the path finder for future. | The operation of the neural networks depends upon the training process. Poorly trained network will operate poorly and outputs cannot be guaranteed. |
| **3.** | Neural networks are able to model very complex functions and detect all possible non linear | Neural Networks are associated with great computational burden. |

 

| | | |
|---|---|---|
| | relationships between inputs and outputs that a linear program cannot. | |
| **4.** | Even if an element of the neural networks fails, it can continue to operate without any problem because of its parallel processing nature i.e. the network is robust and fault tolerant | Neural Networks approach is like black box technique as we are unable to understand the underlying structure of the network. |

## 5. 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] Dilip Kumar Yadav, S.K. Chaturvedi, Ravindra B. Misra, "Early Software Defects prediction using fuzzy logic", International Journal of Performability Engineering, Volume 8 Number 4, July 2012- Paper 6-pp. 399-408.
[2] Gabriela Czibula, Zsuzsanna Marian, Istvan Gergely Czibula, "Software defect prediction using relational association rule mining", Information Sciences, Volume 264 pages 260-278, April 2014.
[3] Jaspreet Kaur, Parvinder S. Sandhu, "A k-means Based Approach for Prediction of Level of Severity of Faults in Software systems", Proceedings of international Conference on Intelligent Computational Systems, 2011
[4] Jun Zheng, " Predicting software reliability with neural network ensembles" , Expert Systems with Applications , Volume 36, Issue 2 , Part 1, March 2009, pp2116-2122.
[5] Karim O.Elish, Mahmoud O. ELish, "Predicting defect prone software modules using Support Vector Machines" Journal of Systems and Software, Volume 81 Issue 5 May 2008, pages 649-660.
[6] Karpagavadivu.K, Maragatham.T, Dr. Karthik.S, "A Survey of Different Software Fault Prediction Using Data Mining Techniques Methods ", International Journal of Advanced Research in Computer Engineering & Technology , Volume 1 Issue 8 , October.
[7] 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.
[8] Lan Guo, Yan Ma, Bojan Cukic, Harshinder Singh, "Robust Prediction of fault proneness by Random Forest", Software Reliability Engineering, 2004 ISSRE 2004.
[9] MaBaojun, Karel Dejaeger, Jan Vanthienem, Bart Baesens, "Software Defect Prediction based on association rule classification" , International journal on Electronics Business Intelligence, pp:396-402
[10] 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.
[11] N.Karunanithi, D.Whitley, Y.K. Malaiya, "Using Neural networks in software reliability prediction" , IEEE Software,Vol.9, no.4, pp. 53-59,1992.
[12] P.K Kapoor, V.S.S Yadavalli, S.K Khatri, M. Basirzadeh. " Enhancing software reliability of a complex software system architecture using artificial neural networks ensemble" , International Journal of Reliabilty, Quality, and Safety Engineering, Vol. 8 Issue 3 , 2011, pp. 271-284.
[13] P.Singh, "Comparing the effectiveness of machine learning algorithms for defect prediction", International Journal of Information Technology and Knowledge management. Volume 2 No. 2, pp 481-483, 2009
[14] 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

 

[15] Rachna Ratra, Navneet singh Randhawa, Parneet Kaur, Dr.Gurdev Singh, "Early Prediction of fault prone modules using clustering based vs. Neural Network Approach in software systems", International Journal of Electronics & Communication Technology, Vol 2 Issue 4 , Oct-Dec 2011

[16] Reena P, Bini Rajan, "Software Defect Prediction System – Decision tree Algorithm with two level Data Preprocessing" , International Journal of Engineering Research & Technology Vol. 3 Issue 3 , 2014.

[17] Ruchika Malhotra, "A Defect Prediction Model for open source software", Proceedings of the worlCongress on Engineering 2012 Vol. 2 WCE, July 4-6, 2012, London U.K.

[18] 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

[19] S.Kanmani, V.R Uthariraj, V.Sankarnarayanam, P.Thambidura, "Object Oriented Software fault prediction using neural networks", Information and Software Technology, Vol. 49 Issue 5 , 483-492, 2007.

[20] Tao WANG, Weihua LI, Haobin SHI, Zun LIU, "Software Defect Prediction on Classifier Ensemble", Journal of Information & Computational Sciences, 8:16(2011) 4241-4254.

[21] T.M. Khoshgaftaar, E.D. Allen, J.P. Hudepohl, S.J. Aud, "Application of neural networks to software quality modelling of a very large telecommunication system", IEEE Transaction on Neural Networks. Vol. 8, No. 4, pp.902-909, 1997.

[22] T.M. Khoshgaftaar, N.Seliya, N.Sundaresh, "An Empirical Study of Predicting software failures with Case Based Reasoning", Software Quality Journal 14(2006)85-111.

[23] Venkata U.B. Challagulla, Farokh B.Bastani, I=Ling Yen, Raymond A. Paul, "Empirical Assessment of Machine Learning based Software defect Prediction Techniques", Proceedings of the 10th IEEE International Workshop on Object Oreinted Real time Dependable Systems, 2005

[24] Xi Tan, Xin Peng, Sen Pan,Wenyun Zhao, "Assessing software quality by program Clustering and Defect Prediction" 18th Working Conference on Reverse Engineering 2011.

[25] 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.

[26] 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

[27] controls.engin.umich.edu/wiki/index.php/NN

[28] www.statsoft.com/Textbook/Data-Mining-Techniques#mining

[29] www.saedsayad.com/data_mining.htm

 