A Survey of Software Defect Prediction using Data Mining Tool

Simpy Awadhiya[1] Dr. Sanjay Agrawal[2]
[1]Student [2]Professor
[1,2]Department of Computer Engineering & Application
[1,2]NITTTR Bhopal

Abstract— In this survey, the authors have discussed the common defect prediction methods utilized in the previous literatures and the way to judge defect prediction performance. Second, we have compared different defect prediction techniques based upon metrics, models, and algorithms. Third, we discussed numerous approaches for cross-project defect prediction that's an actively studied topic in recent years. We have them discuss the applications on defect prediction and alternative rising topics. Finally, we have determined problem areas of the software defect prediction which would lay the foundation for further research in the field.

Key words: Software Defect Prediction, Data Mining

I. INTRODUCTION

It has dependably been extremely crucial to the prescient nature of the software being built. A great deal of strategies has been proposed and is being utilized as a part of various periods of the software development life cycle. Be that as it may, utilization of these methods is obliged by time and spending plan apportioned to extend. Venture supervisors are more than frequently compelled to remove a few stages of the life cycle all on the grounds that to deliver the item on time. Along these lines, any device which may organize the succession of execution of errand in any periods of the development cycle is viewed as exceptionally convenient. This in the long run prompts removing pointless undertakings and execution of organized assignments. One such zone is software testing, a great deal of modules is worked over existing models and have slightest odds of creating bugs, be that as it may, all these models are tried without much thought and subsequently prompts deferred ventures. Subsequently, software imperfection prediction has turned into an essential piece of value administration tools in the present situation.

Software defect – It is any stream or defect in a software work item or software process. To predict these deformities various strategies have been utilized viz. Information mining.

II. SOFTWARE DEFECT PREDICTION

A. Software Defect Prediction Model

Software Defect Prediction Model alludes to those models that attempt to predict potential software defects from test data. There lies an association between's the software metrics and fault-proneness of the software. A Software defect prediction models comprise of independent variables (Software metrics) gathered and measured software development life cycle and dependent variable. There are different data mining techniques for defect prediction.

B. Data Mining Analysis

Data mining is the investigation venture of the "Knowledge Discovery in Databases" process, or KDD, a process of finding examples in substantial data sets including strategies at the crossing point of artificial intelligence, machine learning, measurements, and database systems. The general objective of the data mining process is to concentrate data from dataset and change it into a sensible structure for further analysis. Data Mining can be separated into two tasks: Predictive tasks and descriptive tasks. The predictive tasks to predicting the estimation of a particular characteristic (target/dependent variable) based on the estimation of different qualities (explanatory). The descriptive task is to infer designs (connection, patterns, and directions) that condense the hidden relationship between data.

C. Data Mining Techniques for Defect Prediction Models

In this paper, diverse data mining methods are discussed for recognizing faults prone modules. Data Mining assumes a critical part in software defect prediction. It helps in cleaning data. For this the data is taken from the Software repositories. It has heaps of data that is helpful in evaluating software quality. Data mining procedures and can be connected on these archives to extricate the valuable data. Data mining strategies can be connected on the product repositories to extricate the defect of a software product. Different data mining procedures utilized for software defect predictions:

D. Clustering

Clustering is an unsupervised learning in which no class marks are given. Clustering is an approach to sort an accumulation of things into cluster or groups whose individuals are comparative someway. It the assignment of the collection an arrangement of things in a manner that things in the same cluster are like each other and not at all like those in different groups [1, 2,3].

1) Software Defect Prediction using Clustering

In [4], they they have utilized k-mean strategy of clustering for software defect prediction. K-mean clustering is a non-hierarchical clustering technique in which things are moved among sets of clustering until the wanted set is coming to. It has certain drawbacks, so to defeat those disadvantages Quad Tree-based k-mean grouping strategy was proposed. The goal was: to begin with, 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 assessed the viability of Quad tree-based k-mean clustering algorithm in predicting faulty software modules. When contrasted with the first k-mean calculation. The consequence of this Quad tree-based k-mean calculation is contrasted and different methodologies what's more, found that the amount of emphases s of k-means computation is less in case of Quad tree-based k-mean aside from different methodologies, and in addition percent blunder likewise give genuinely satisfactory qualities.

E. Classification

Classification is a procedure of finding an arrangement of models that describe and distinguish data classes or concepts Number conditions sequentially. It comprises of predicting a specific result taking into account a given information. The Classification method use input data, additionally called preparing set where all items are as of now labeled with known class names. The goal of classification algorithm is to break down and gains from the preparation training data set develop a model. This model is then used to classify test data for which the class names are not known. [1, 3,5 6]. The different order strategies are given below.

- Neural Networks: Neural Networks are the non linear predictive models which can learn through training and look like natural neural systems in structure. A neural traditions comprises of joint control components assembled neurons that work in parallel inside a system to create yield. [6, 7].
- Decision Trees: A Decision tree is a prescient model which can be utilized to represent to both classification and regression models in the structure a tree structure. It refers to a various hierarchical model of decision and their outcomes. It is a tree with decision nodes and leaf node. A decision node has two or more branches. Leaf nodes represent a classification or decision. [8,9].
- Naye Bayes: It is based on Bayes Classifier depends on the supposition that the nearness or nonappearance of a specific element of a class in not identified with the nearness or nonattendance of whatever other component. [21, 10].
- Support Vector Machines: SVM depends on the idea of decision planes that define decision boundaries. A decision plane is the one that isolates between arrangements of items having a diverse class enrollment. SVM is principally a classifier technique that performs grouping assignment by building hyper plane in a multidimensional space that isolates instances of various class names. It supports both regression and classification. [10, 23].

F. Association

The Association mining undertaking involves perceiving the ceaseless item sets, and a short time later implication rules among them. It is the errand of finding connections between things in information sets. It is a procedure for finding charming associations between variables in vast databases. It is about finding association or correlations among sets of items or objects in database. It basically oversees finding concludes that will anticipate the event of thing in perspective of the events of various things. [13, 14].

1) Software Defect prediction using association mining

In association rule mining procedure we use defect type data to predict software defect associations that are the relations among various defect types. The defect associations can be utilized for three purposes: First, Find whatever number of related defects as possible to the detected defects and make more successful revisions to the software. Second, it assesses analyst's outcomes amid an inspection. Third, it helps in helping supervisors in enhancing the software process through analysis of the reasons some defect every now and again happen together. Association rule mining goes for finding the examples of co-events of the characteristics in the database. They found that higher confidence levels may not result in higher prediction accuracy.

G. Regression

It is a statistical process to evaluate the relationship among variables. It examinations the relationship between the dependent or response variable and independent or predictor variables. The relationship is as a condition that predicts the response variable as an immediate limit of indicator variable. [16,18]. Information can be smoothed by fitting the information to a capacity, for example, with relapse straight relapse includes finding the "best" line to fit two variables are included and information are fit to the multidimensional surface utilizing relapse to discover scientific condition to fit the information smooth out the commotion. Linear Regression: eY=a+bX+u

III. Software Defect Prediction (SDP) using Different Classification Techniques

A study is directed to help developers identify defects in view of existing software metrics utilizing data mining methods particularly Classification and there by enhance software quality which prompts diminishment in the software development cost in the improvement and maintenance stage. Diverse ordering procedures have been overviewed with different data sets.

A. SDP using Supervised Learning

The various Supervised Learning techniques are discussed in this section.

1) SDP using Bayesian Network

Yuan Chen, et.al [19] has over viewed the diverse data mining, classification techniques for software defect prediction. They proposed another model taking into account Bayesian network and PRM to predict the software defect and mange.

Thair Nu Phyu [15] reviewed on various classification techniques, for example decision tree prompting, Bayesian networks, k-nearest neighbor classifier, case-based reasoning,

Thair Nu Phyu [15] reviewed on various classification techniques such as decision tree induction, Bayesian networks, k-nearest neighbor, case-based reasoning, genetic algorithm and fuzzy logic method. K-nearest neighbor, case-based reasoning, genetic algorithm and fuzzy logic method. The results found that there is no proper info that which is the best classifier. Several of the classification produce a set of interacting loci that best predict the phenotype. In any case, a direct use of characterization strategies to substantial quantities of markers has a potential risk picking up randomly associated markers.

Wen Zhang et.al [19] proposed Bayesian Regression Desire Maximize calculation for software effort prediction and two embedded methodologies handle missing data. They utilized the technique for disregarding the missing data in an iterative way in the predictive model. Here they have utilized data sets, for example, ISBSG and CSBSG. At the point when there are no missing data BREM with CR, BR, and SVR& M5. At the point when there are missing data BREM with MDT and MDI beats attribution

method incorporates MI, BMI,CMI, and Mini& M5. BRM is used for software prediction and MDI used for finding missing values embedded with BREM.

Arvinder Kaur and InderpreetKaur[20], they have attempted to quality of the software product based on identifying the defects in the classes. They have done this by utilizing six unique classifiers, for example, Naive base, Logistic relapse, Instance based (Nearest-Neighbor), Bagging, J48, Decision Tree, Random Forest. This model is connected on five diverse open source software to discover the defect of 5885 classes based on object oriented metrics. Out of which they observed just Bagging and J48 to be the best.

K.Sankar et.al [11], proposed a system which conquers the issue of deficiency in the processing precision and utilization of the huge number of components. This paper proposed Feature determination method Feature selection technique to predict faults in software code and it also measure the software code and performance of Naive based and SVM classifier. The precision is measured by F-mean metric. Hereditary calculation and fuzzy logic techniques. The outcomes found that there are no appropriate data that which is the best classifier. A few of the other techniques create an arrangement of collaborating loci that best anticipate the phenotype. Be that as it may, a clear use of arrangement techniques to extensive quantities of markers has a potential danger getting randomly related markers.

2) SDP using Ensemble Method/ Random Forests

Issam Het.al[22] have proposed a two-variation ensemble learning classifier which demonstrates that the greedy forward choice is superior to anything relationship forward selection. Further, they proposed a model APE with seven unique classifiers which comes about much better when contrasted with weighted SVM's and random forests. Further, they upgraded the variant of APE with insatiable forward determination to deliver higher AUC measures for the distinctive information sets. The outcomes indicated more redundant and irrelevant features.

Renqing Li &Shihai Wang [23] predicted defects imbalanced information sets. C4.5, SVM, KNN, Logistic, Naïve Base, Ada help &smooth support models were tried on imbalanced data sets of NASA's MDP. The outcomes found that Smooth help observed to be the best best defect predictor when compared to the others.

C.Chung and S. Dhall [25] proposed different classification methods to predict software defect. Here Three sorts of a classifier, for example, J48, Random Forest and Naive Bayesian Classifier is connected to different on going data sets of NASA to assess the data sets taking into account different criteria like ROC, Precision, MAE, RAE etc.

3) SDP using Support Vector

Machine Sonali Agarwal and DivyaTomar[24] have propositioned a part decision based Linear Twin Support Vector Machine (LSTSVM) model predict defect prone inclined software modules. F-score strategy is utilized for software defect expectation taking into account different software metrics. This model is connected to PROMISE information sets and contrasted and the other existing models. The outcomes say that the execution of the new model is superior to the current machine learning models.

CagatayCatal[24], proposed four semi-supervised classification techniques, for example, Low- density separation (LDS), support vector machine (SVM), (EM-SEMI), and class mass normalization class mass standardization (CMN) for semi-administered defect prediction. They associated 4 sorts of ssc on NASA datasets. The results exhibited that SVM and LDS are superior to anything CMN and EM-SEMI. LDS performs much superior to anything SVM for an extensive data set.

David Grayet.al proposed a work utilizing the static code measurements for an accumulation of modules contained inside eleven NASA data sets are utilized with a Support Vector Machine classifier. A thorough grouping of pre-preparing steps was connected to the information before characterization, including the adjusting of both classes (defective or something else) and the evacuation of countless occurrences. The Support Vector Machine in this test yields a normal precision of 70% of beforehand concealed data.

4) SDP using Decision Tree

Thair Nu Phyu [15] checked on different classification techniques, for example, decision tree instigation, Bayesian network, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques.. The outcomes found that there is no appropriate information that which is the best classifier. A few of the classification techniques create an arrangement of interfacing loci that best predict the phenotype. However, a direct use classification strategies to expansive quantities of markers has a potential danger getting randomly associated.

B. SDP using Semi-Supervised Learning

Ming Ming Li, et al.proposed [27] a sample based methods for software defect prediction. Three strategies, for example, random sampling with traditional machine learners, random sampling with a semi-supervised learner as well as active sampling with active semi-supervised learning. They connected a semi-regulated learning technique called ACoForest to create a classification tool for the remaining un-sampled modules. They likewise proposed a novel dynamic semi administered strategy called AcoForest which can choose unsampled modules and investigated Promise data sets and observed to be the best technique. Trial results demonstrate that size does not influence the defect prediction.

CagatayCatal[24]proposed four sem administered arrangement techniques, for example Low-density separation (LDS), support vector machine (SVM), expectation-maximization (EM-SEMI), class mass normalization (CMN) for semisupervised defect prediction. They connected 4 sorts of ssc on NASA datasets. The outcomes demonstrated that SVM and LDS are superior to anything CMN and EM-SEMI. LDS performs much superior to anything SVM for a large data set.

C. SDP using Unsupervised Learning

C.Chung and S.Dha[25] proposed a different classification and clustering methods to predict software defect. The various data mining classifier algorithms specific J48, Random Forest, and Naive Bayesian Classifier (NBC) are assessed taking into account different criteria like ROC, Precision, MAE, RAE etc.

Clustering technique is later applied to different data set of NASA working k-means, Hierarchical Clustering and proposed a different grouping and Density Based Clustering algorithm to predict software defects. The different data mining classifier calculations to be specific J48, Random Forest, and Naive Bayesian Classifier (NBC) are assessed taking into account different criteria like etc.

D. SDP using Machine Learning Algorithm

Xiao-YuanFing, et.al [26] have attempted to show the powerful, productive and low computational weight utilizing propelled machine learning method, for example, cooperative delegate arrangement. The new model proposed by them is CSDP which is utilized to be used to predict defect in a very efficient manner.

IV. CONCLUSION AND FUTURE WORK

In this paper we have discussed how various data mining technologies like-Decision tree, semi-supervised learning, unsupervised learning and support vector- have been used for calculating defect probability. We made an extensive study of these methods and accessed their performances on various data sets for defect prediction. After detailed survey following shortcoming was concluded in these papers.

1) The classifiers being utilized as a part of past work have been superseded by more efficient classifiers.
2) Most of the tests in reviewed papers have been completed utilizing PROMISE repository, which as of now contains pre-processed data, consequently the vast majority of these works don't put emphasis on change in pre-processing techniques.

In our future works we have planned to use NP-SVM as classifier as it outperforms other existing classifiers which have been used so far.

REFERENCES

[1] Balaji, V.Venkateswara Rao, "Improved Classification Based Association Rule Mining", International Journal of Advanced Research in Computer and communication Engineering, vol. 2, no. 5, (2013).
[2] D.Mehta, "A Comparative study of Techniques in Data Mining", by Manika Verma1, International Journal of Emerging Technology and Advanced Engineering, vol. 4, no. 4, (2014).
[3] A. Chug1 and S. Dhall1, "Software Defect Prediction using Supervised Learning Algorithm and Unsupervised Learning Algorithm", The Next Generation Information Technology Summit (4th International Conference), (2013), pp.1-6.
[4] Partha Sarathi Bishnu and Vandana Bhattacherjee, "Software Fault Prediction Using Quad Tree-Based KMeans Clustering Algorithm", IEEE Transactions on knowledge and data engineering, Vol. 24, no. 6, June 2012.)
[5] V. Ajay Prakash, D. V. Ashoka, V. N. ManjunathAradya, "Application of Data Mining Techniques for Defect Detection and Classification", Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, Advances in Intelligent Systems and Computing, vol. 327, (2015), pp. 387-395.
[6] A.TosunMisirli, A. se Ba¸ S.Bener,"A Mapping Study on Bayesian Networks for Software Quality Prediction", Proceedings of the 3rd International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, (2014).
[7] C. Catal, "A Comparison of Semi-Supervised Classification Approaches for Software Defect Prediction", Journal of Intelligent Systems, vol. 23, no. 1, pp. 75-82,(2013).
[8] "Software defect prediction using supervised learning algorithm and unsupervised learning algorithm", Confluence 2013: The Next Generation Information Technology Summit (4th International Conference), (2013).
[9] S. Kaur, and D. Kumar, "Software Fault Prediction in Object Oriented Software Systems Using Density Based Clustering Approach", International Journal of Research in Engineering and Technology (IJRET) vol. 1, no. 2,(2012).
[10] L. Li, H. Leung, "Bayesian Prediction of Fault-Proneness of AgileDeveloped Object-Oriented System:Lecture Notes", Business Information Processing, vol. 190, (2014), pp. 209-225
[11] K. Sankar, S. Kannan and P.Jennifer, "Prediction of Code Fault Using Naive Bayes and SVM Classifiers Middle-East Journal of Scientific Research", vol. 20, no. 1, (2014), pp.108-113.
[12] R. M. Rahman, F. Afroz,"Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis",Journal of Software Engineering and Applications, (2013), vol.6, pp.85-97
[13] G.Czibula, Z. Marian, I. G.Czibula, "Software defect prediction using relational association rule mining, Information Sciences", vol. 264, no. 20 (2014), pp. 260-278.
[14] 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).
[15] T. Nu Phyu, "Survey of Classification Techniques in DataMining", International MultiConference of Engineers and Computer Scientists, (2009); Hong Kong.
[16] R.Kalsoom, M. Qureshi, "Application and Verification of Algorithm Learning Based Neural Network",arXiv preprint arXiv:1406.2614, (2014), arxiv.org.
[17] 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),.
[18] C. Catal, U.Sevim, B. Diri,"Practical development of an Eclipse-based software fault prediction tool using Naive Bayes algorithm", Elsevier,(2011).
[19] W. Zhang, Y. Yang, Q. Wang, "Using Bayesian Regression and EM algorithm with missing handling for software effort prediction", Information and software technology, vol. 58, (2015), pp. 58-70.
[20] A. Kaur and I. Kaur, "Empirical Evaluation of Machine Learning Algorithms for Fault Prediction", Lecture Notes on Software Engineering, vol. 2, no. 2, (2014).

[21] I. H. Laradji, M.Alshayeb, L.Ghouti, "Software defect prediction using ensemble learning on selected features. Information and Science Technology", vol. 58, (2015), pp. 388-402.
[22] R. Li, S.Wang, "Ann Empirical Study for Software Fault-Proneness Prediction with Ensemble Learning Models on Imbalanced Data Sets", Journal of Software, vol. 9, no.3,pp. 697-704,(2014).
[23] S. Agarwal and D.Tomar, "A Feature Selection Based Model for Software Defect Prediction", International Journal of Advanced Science and Technology, vol.65,(2014), pp. 39-58.
[24] C. Catal, "A Comparison of Semi-Supervised Classification Approaches for Software Defect Prediction", Journal of Intelligent Systems, vol. 23, no. 1, pp. 75-82,(2013).
[25] A. Chug1 and S. Dhall1, "Software Defect Prediction Using Supervised Learning Algorithm and Unsupervised Learning Algorithm",The Next Generation Information Technology Summit (4th International Conference),(2013),pp.1-6.
[26] P.Dhiman, M.C. Manish,"A Clustered Approach to Analyze the Software Quality Using Software Defects, Advanced Computing & Communication Technologies (ACCT)", 2012 Second International Conference,(2012).
[27] M. L., H. Zhang, R. Wu, Z.-H. Zhou, "Sample-based software defect prediction with active and semisupervised learning", Automated Software Engineering , (2012), vol. 19, no. 2, pp. 201-230

