ISSN 2347 - 3983
Volume 8. No. 6, June 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter62862020.pdf
https://doi.org/10.30534/ijeter/2020/62862020

Software Defect Prediction Utilizing Deterministic and Probabilistic Approach for Optimizing Performance through Defect Association Learning

Rohini B. Jadhav[1],Shashank D. Joshi2,Umesh G. Thorat3,Aditi S. Joshi[4]
[1]Research Scholar, Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune, India, rohini.jadhav@outlook.com
[2]Professor, Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune, India,sdj@live.in
[3]Solution Architect, Tata Consultancy Services Pune, India,umesh.thorat@tcs.com
[4]Research Scholar, Pune Institute of Computer Technology Pune, India,aditijoshi14@gmail.com

ABSTRACT

While developing software it is very important that the software should be of defect free. But, none of the software can be 100% defect free and various studies are in progress to build a model which minimizes the defect as much as possible by predicting it at an early stage of development. Based on the probability facts various researchers has used probabilistic model to predict defects in the program. To contribute in this research and enhancing the existing model of software defect prediction we are proposing a model based on the combination of probabilistic and deterministic model through defect association learning. The experimental evaluation in comparison with the existing methods shows the improvement in the accuracy of predicting the defect by using Deterministic and Probabilistic defect prediction.

Key words: Association Learning, Software Defect Prediction, Probabilistic, Deterministic,

1. INTRODUCTION

Software program is one of the crucial aspects of a system in a domain. An ideal software system should improve its vital innovations and functionality and it is important to develop programs with high quality for secure and reliable systems. But, to provide a quality and reliable software it leads to more time and efforts in various process like validation, verification etc. Therefore, to measure the effort and time required for software engineering and defect prediction is very difficult [1] [2].

One of the popular and most accepted study of data mining is Association Rule based mining [3]. This technique mainly focuses on identifying the important correlation between records [4]. There are many tasks that are important in data mining technique and one of those are to construct a fast and accurate classifier for defective records [5]. Existing association rules mostly based on classifiers having higher accuracy in classification and also based on objects set's confidence and support rules. So, there should be some changes that need to be made such that the recorded defect prediction should have appropriate classification for predicting the defect.

Many researchers investigated and identified that the use of deterministic models [6], [7] in predicting the unambiguous or results having exact match is effective. There are various real-time applications such as "Google", "Twitter" and "Face book" in which users can be easily assigned deterministically for predictions and analysis. In this model, considerable aspects are utilized for predicting a result which is much closed to the precise value while in contrary; these aspects are used to predict a similar result in a probability distribution i.e. in non deterministic model.

Whenever any defect prediction model is proposed it is very important to measure the performance of that model depending on the whether the ignored elements is not really vital to investigate in the underlying phenomenon. As well as it is very difficult to confirm that a particular approach or mathematical model is appropriate before some observations are tested. So to confirm whether a particular model is accurate enough to conduct defect prediction we should first test the model under different condition through actual observations. To enhance the real time defect prediction, it is necessary to understand defective attributes in the system with certain determinism decision. In this paper, we present a model in which defect prediction is done by applying the concept of Deterministic and Probabilistic approach through Defect Association Learning.

The latter part of the paper is organized into 5 sections. In the section-2 the Literature survey in relevance to the proposal was

discussed, in section-3 the objectives are identified and presented in section-4 the proposed DP-DP methodology and Defect Association Learning process is presented, in section-5 the experimental evaluation and the comparison of results analysis comparison is presented, and in section-5 it presents the conclusion.

2. LITERATURE SURVEY

There are various researches carried out in the domain of software defect prediction where number of researchers has investigated and identified different techniques to achieve maximum accuracy for predicting the defect .There are many software technologies [8], [9] to support "log-based defect analysis" have been developed till date and for the modelling and processing of historical data, such as, " Analyze NOW" [10], "MEADEP" [11] etc. the modern capture techniques are integrated. But it does not come up by practices that are completely automated. There are certain pre treatment activity to analyze the defects[12] but in many cases only few studies are successful to reduce test effort and have quality software by detecting the defect in early stage of software development.[13][14][15][16].

T. Mended ET. [17] Suggested that whatever the effort done should be measured so that the defect prediction accuracy can be accessed. As well as he said that the various traditional metrics selected for predicting the defect such as "accuracy", "recall", and "ROC curves" ignore the quality assurance cost .Another researcher C. F. Kemmerer et al. [18] examined how the software quality gets impacted by the the effect of the verification rate, by investigating the inspection aspect of verification techniques.

In a probabilistic or non-deterministic model, the experiment is observed under certain scenarios which will only determine the probabilistic behaviour of observed result. For instance, suppose we would like to predict the amount of rainfall due to a particular storm passing through a specific location... Meteorology might provide essential data about atmospheric pressure at different points, changes in pressure[19], the nearby storm system, origin and direction of the storm, and so on. Tools are available which are required to record precipitation gaining all this information will not give the accurate data related to rainfall but rather it will give probable value more precisely. Physical considerations are used by deterministic models to predict accurate end result while more probability distribution can be predicted by probabilistic model using these considerations.[19][20].

Suppose that we have a software having 4 modules out of these 4 modules one module are faulty as the software does not work as desired. If the faulty module is not corrected by the system then it will result loss in time and cost. So avoid this, all modules must rest to identify the actual defect. If we assume that prediction model has the knowledge of the defect occurred as it is has handled similar scenario in the past then the prediction likelihood may be true or false. It is very important and compulsory to make prediction to overcome such probabilistic methods. Many probability models has 90% accurate prediction rate but to get more accurate result it needs to be correlated the defect that has relevant and specific data.

3. OBJECTIVE

Keeping the research indications in view, it is realized that there exists enough scope to improve the software defect prediction. The objectives are confined to the following:

1. To use better evaluation measurement parameter to get better result
2. To decrease the software development cost, time and effort
3. To propose a model that accurately predict the software defect
4. To implement a method using deterministic approach this will help optimising the result provided by probabilistic model.
5. To propose a hybrid model for defect prediction based probabilistic and deterministic approach for better performance.

4. PROPOSED METHODOLOGY

By identifying the various limitations in the existing system and formalizing the objectives, following is the proposed architecture of Deterministic and probabilistic defect prediction through Defect Association learning

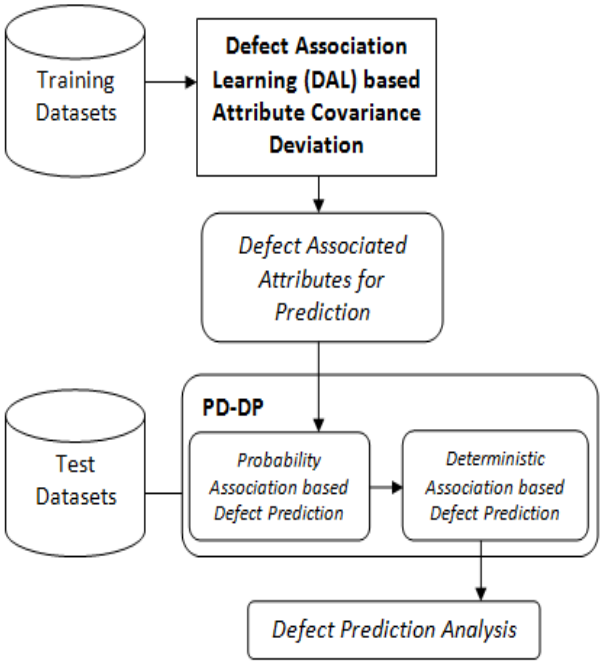

Figure.1: System Architecture for Defect Prediction Mechanism

The architecture is mainly divided into 3 main modules
A. Defect Association Learning
B. Probabilistic Association Method
C. Deterministic Association Method

A. Defect Association Learning
Association rule mining is a data mining techniques which used to find out the correlation between different attributes by generating the patterns so that it is easier for prediction, classification or decision making. Defect Association Learning is based on association rule mining in which association among defect is very important to build associated rules for predicting the defects accurately.

Defect Association Learning is one of the pillars in this research to accurately predict defect. So in DAL we can build associated rules by using various attribute patterns. For effective predicting the defect and decision making DAL uses various datasets having multiple defect set from which it combines high impact defect associated pattern.

To predict the association rules DAL implements 2 mechanisms. It initially uses attribute reduction mechanism through which it can find the association between the attributes and attributes patterns are generated, which are combined so that the efficient defect prediction rules is formed.

The highly important attribute for defect prediction can be found by using the attribute reduction mechanism. Covariance Deviation measure is used by DAL for measuring the association between each attribute so the attribute having high impact on defects can be found.

Suppose we have two attributes P and Q having the n values an) which is calculated using the following equations Eq. (1) and (2)

$$En(P) = \bar{P} = \frac{\sum_{v=1}^{n} p_v}{n} \quad (1)$$

$$En(Q) = \bar{Q} = \frac{\sum_{v=1}^{n} q_v}{n} \quad (2)$$

Based on the computed entropy value *En* of the attribute we will now compute each attributes Covariance Deviation using Eq. (3) as,

$$CD_A = \sum_{v=i+1}^{n} H(P) - H(Q_v) \quad (3)$$

After computing the Covariance Deviation value from above equation (3) it creates a attribute set as A which has values > = 1. The variance value > = 1 indicates that it have high impact on prediction and if the value < 1 it indicates that it have low impact on prediction of defects.

Now, we get a set of attributes names *A* from the associated and reduced attributes, it will build the defect associated attributes which is required for prediction of each defect. This will provide efficient prediction rules and reduces the computational overhead.

B. **Probabilistic and Deterministic based Defect Prediction**

Association Rule Mining is one the techniques on which many of the existing prediction classifiers [21] are based on. The rules obtained are labeled and trained. For classification these rule are learned from dataset and applied on another data instance. So in the same way the proposed method of Defect Prediction based on Deterministic and Probabilistic facts utilize the trained and labeled data obtained from Defect association learning in the form of reduced attribute set as A, and its unique combination of items are used to generate each attribute pattern to perform defect prediction. The proposed method of defect prediction includes two methods as, a) Probabilistic Association Method and b) Deterministic Association Method. The details of these methods are given below

Probabilistic Association Method

We have a set *A* having reduced attributes from which a pattern *T* is generated based on the attribute value for defect prediction. Let's assume that we have a defect dataset as $S_{n..}$ For each attribute it generates a different and individual pattern by using a Defect Association Learning method using Eq. 4) as,

$$T_n = R(A_k) \quad (4)$$

where "*R*" → item set extraction data processing method "T_n " → Each attribute extracted pattern and n=*1, ... , N,* and "A_k " → Defect attributes value from *k=1, ... , K.*

Each attribute's as A_k obtained patterns as T_n for each dataset of S_n will merge to generate a combined pattern as T_k

$$T_k = C_A(T) \quad (5)$$

Using, the equation (4) and (5), for all dataset, S_n a new defect prediction rule will be formed as,

$$T_n = R\ (A_1 \wedge \dots\dots \wedge A_k) \longrightarrow F_k \quad (6)$$

$$P := C\ (P_1 \wedge \dots\dots \wedge P_n) \longrightarrow Q \quad (7)$$

where," F_k " → Associated Pattern, and " Q " → Qualified patterns for the defect prediction rules.

Deterministic Association Method

This method is used to increase the accuracy of the prediction model . It utilizes the associated defected records which are generated by probability association method to perform defective data prediction. Now, suppose we have a test data set which is predicted as defective by probability association method as Y_k *which is* having attributes A_x .To get the combined pattern we combine Eq.(4) and Eq. (5) to get input data records pattern which is presented in Eq. (8) and Eq. (9)

$$T_n = R\ (A_1, \dots \dots, A_x) \quad (8)$$

$$T_k = C_F(P_n) \quad (9)$$

Now, to have accurate prediction using DAM for each predicted records as E the correlation of the pattern is obtained through probability association method by measuring "Support", "Confidence" and "Lift" measures using the following Eq.(10),(11)and (12)

$$Support = Prob(E \wedge Q) \quad (10)$$

$$Confidence = Prob(E \wedge Q) / (Prob(E)) \quad (11)$$

$$Lift = Prob(E \wedge Q) / (Prob(E) * Prob(Q)) \quad (12)$$

The Lift measure using Eq. (12)provides the prediction deterministic. If the lift value is greater or equal to one with the *Q* rules patterns generated by DAL, then the prediction is considered to be positive and we get a qualified defect prediction, and if the value is less than one then the prediction is considered as negative.

5. EXPERIMENTAL EVALUATION

The dataset required for evaluating the proposed approach is NASA PROMISE Repository[22][23] of the dataset KC3,PCI,PC2,JM1,and CM1. All the mentioned datasets have different attributes as well as defect ratios and modules. Following Table 1 is the dataset description.

Table 1: Description of Dataset

Modules	Number of Instances	% of defects
PC1	759	8.1 %
PC2	1585	1.0 %
CM1	327	12.8 %
KC3	200	18.0 %
JM1	9371	18.5 %

The records in the dataset with defect value or class value given as "true" or "false. If the class value is "true" then it indicates that it is reported defects whereas if class value is "false" then it indicates that there may be or not a defect in the module.

The above mentioned datasets are demonstrated in contrast with the existing classifier which are based on probabilistic approach such as "One R", "BayesNet", "JRip" etc. The Defect Association Learning is implemented using java and WEKA tool is used to evaluate performance of the proposed module.

Defect Association learning process results in attribute pattern selection. On the basis of that we implement our proposed mechanism Deterministic and Probabilistic Defect Prediction to measure the accuracy of defect prediction and also to compare it with other existing techniques and evaluate the performance of the model . Below Table-2 shows the outcome of the proposed model in comparison with other existing models.

Table 2: Defect Prediction Accuracy and Relative Abs. Error Comparison

Classifiers	Modules	Correctly Classified	Incorrectly Classified	Prediction Accuracy	Relative Absolute Error
BayesNet	PC1	701	58	92.358	7.642
	PC2	1351	234	85.237	14.763
	CM1	298	29	91.131	8.869
	KC3	185	15	92.500	7.500
	JM1	7681	1690	81.966	18.034
NaiveBayes	PC1	721	38	94.993	5.007
	PC2	1453	132	91.672	8.328
	CM1	304	23	92.966	7.034
	KC3	186	14	93.000	7.000
	JM1	8457	914	90.247	9.753
JRip	PC1	685	74	90.250	9.750
	PC2	1247	338	78.675	21.325
	CM1	285	42	87.156	12.844
	KC3	181	19	90.500	9.500
	JM1	7541	1830	80.472	19.528
OneR	PC1	698	61	91.963	8.037
	PC2	1326	259	83.659	16.341
	CM1	288	39	88.073	11.927
	KC3	176	24	88.000	12.000
	JM1	7210	2161	76.939	23.061
PD-DP	PC1	748	11	98.551	1.449
	PC2	1504	81	94.890	5.110
	CM1	315	12	96.330	3.670
	KC3	198	2	99.000	1.000
	JM1	9018	353	96.233	3.767

We have compared the different models based on the accuracy and relative absolute error . In Figure. 2 we are comparing defect prediction accuracy of different classifiers with our classifier and it indicates that our model achieves 6% higher accuracy than existing classifier and it also has low defect rate. This improvement is because attribute selection by defect association learning is precise and it also

predicted the defected classes with 2% confidence, support and lift measures (Discriminated Methods).

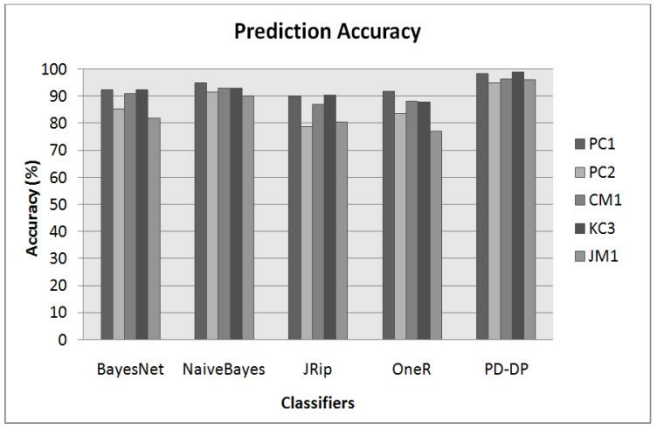

Figure.2: Defect Prediction Accuracy

In Figure.3 we are comparing the relative absolute defect which indicates that the proposed model have a low defect rate as compared to that of the existing one. It is also observed that NaiveBayes showing better accuracy in defect prediction but the average is low by 3% as compared to our model . Both use probabilistic facts but due to integration of deterministic fact our model is achieving more average accuracy in comparison. As well as it is also observed that the other three models i.e. OneR, JRip, BayesNet are having similar error rate as well as accuracy.

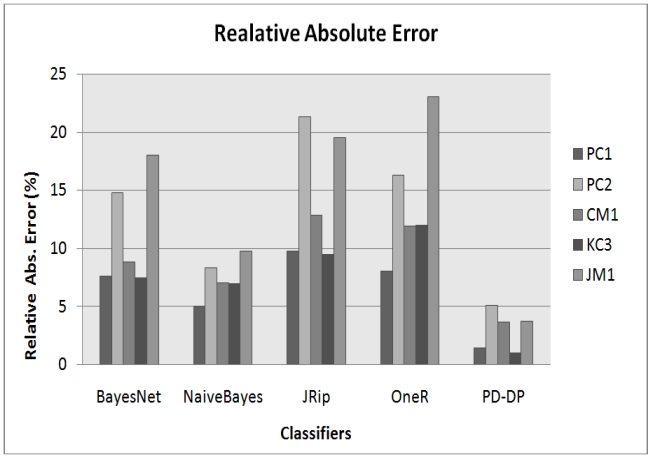

Figure.3 Relative Abs. Error Rate Comparison

6. CONCLUSION

This research implements a Defect Prediction approach based on Deterministic and Probabilistic facts for achieving better accuracy . But before actually applying the deterministic and probabilistic approach it undergoes a Defect Association learning process in which the defect attributes are associated in such a way that the Covariance Deviation can be measured between the various attributes for finding the most impacting attributes for prediction of defect .This model implements 2 methods for achieving the better accuracy namely Probability Association Method (PAM) and Deterministic Association Method (DAM). In PAM, we associate the defect pattern through DAL to the probability of the test data which results in classification of defects as defective or non defective. Further we apply the DAM so that the defect can be predicted more accurately while developing the software. It computes the accuracy using lift , confidence and support measures. We have also conducted the experimental evaluation which shows that the proposed methodology not only improves the accuracy but also decreases the relative absolute error in comparison with the existing defect prediction approaches. In future this approach can also work in analysis of critical defect and decision making.

REFERENCES

1. H. Liang, Y. Yu, L. Jiang, Z. Xie, "Seml: A Semantic LSTM Model for Software Defect Prediction", IEEE Access, Volume: 7, 2019. https://doi.org/10.1109/ACCESS.2019.2925313
2. Y. Kamei, E. Shihab, B. Adams, A. E. Hassan, A. Mockus, A. Sinha, and N. Ubayashi, "A Large-Scale Empirical Study of Just-in-Time Quality Assurance", IEEE Transactions On Software Engineering, Vol. 39, No. 6, June 2013 https://doi.org/10.1109/TSE.2012.70.
3. X. Xue, C. Yao, W. Yan-en, "Study on Mining Theories of Association Rules and Its Application" International Conference on Innovative Computing and communication, 978-0-7695-3942-3/10, 2010.
4. Q. Song, M. Shepperd, M. Cartwright, C. Mair, "Software Defect Association Mining and Defect Correction Effort Prediction", IEEE Transactions on software engineering, Vol. 32, no. 2, 2016.
5. J. -h. REN, F. LIU, "Predicting Software Defects Using Self-Organizing Data Mining", IEEE Access, 2019.
6. W. Charlotte Werndl, "On the observational equivalence of continuous-time deterministic and indeterministic descriptions", European journal for philosophy of science, Vol. 1(2), pp. 193-225, 2011. https://doi.org/10.1007/s13194-010-0011-5
7. C. Woodruff, L. Vu, K. A. Morgansen, D. Tomlin, "Deterministic Modeling and Evaluation of Decision-Making Dynamics in Sequential Two-Alternative Forced Choice Tasks", IEEE Proceedings, Vol.100 (3), 2012.
8. J. Chen, Y. Yang, K. Hu, Q. Xuan, Y. Liu, C. Yang, "Multiview Transfer Learning for Software Defect Prediction", IEEE Access, Vol. 7, 2019.
9. C. Tantithamthavorn, S. McIntosh, A. E. Hassan, K. Matsumoto, "An Empirical Comparison of Model Validation Techniques for Defect Prediction Models", IEEE Transactions on Software Engineering, Volume 43, No. 1, pp. 1 - 18, 2017. https://doi.org/10.1109/TSE.2016.2584050
10. D. Tang, M. Hecht, J. Miller, and J. Handal, "Meadep: A Dependability Evaluation Tool for Engineers",

IEEE Transaction Reliability, Vol. 47(4), pp. 443-450, 1998.

11. Thakur and R.K. Iyer, "Analyze-Now—An Environment for Collection and Analysis of Failures in a Networked of Workstations", IEEE Transaction Reliability, Vol. 45(4), pp. 561-570, 1996.
12. D. -L. Miholca, "An Improved Approach to Software Defect Prediction using a Hybrid Machine Learning Model", IEEE 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2018.
13. E. A. Felix, S. P. Lee, "Integrated Approach to Software Defect Prediction", IEEE Access, Volume 5, pp. 21524 - 21547, 2017.
14. X. Yu, J. Liu, Z. Yang, X. Jia, Q. Ling, S. Ye, "Learning from Imbalanced Data for Predicting the Number of Software Defects", IEEE 28th International Symposium on Software Reliability Engineering (ISSRE), pp. 78 - 89, 2017.
15. M. A. Kabir, J. W. Keung, K. E. Benniny, M. Zhang, "Assessing the Significant Impact of Concept Drift in Software Defect Prediction", IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Vol. 1, 2019. https://doi.org/10.1109/COMPSAC.2019.00017
16. Molawade, Mayuri. (2019). Software reliability prediction using Knowledge Engineering approach. International Journal of Advanced Trends in Computer Science and Engineering. 8. 2768-2772. 10.30534/ijatcse/2019/14862019.
17. T. Mende and R. Koschke, "Effort-Aware Defect Prediction Models", Proc. European Conf. Software Maintenance and Reeng., pp. 109-118, 2010.
18. C. F. Kemerer and Mark C. Paulk, "The Impact of Design and Code Reviews on Software Quality: An Empirical Study Based on PSP Data", IEEE Transactions on Software Engineering, Vol. 35(4), 2009.
19. R.B.Jadhav, S.D.Joshi, Umesh Thorat, "A Probabilistic and Deterministic based Defect Prediction through Defect Association Learning in Software Development", IJRTE Volume-8 Issue-6, March 2020.
20. Jadhav, Rohini. (2019). A Software Defect Learning and Analysis Utilizing Regression Method for Quality Software Development. International Journal of Advanced Trends in Computer Science and Engineering.12751282.10.30534/ijatcse/2019/38842019.
21. E. Baralis, L. Cagliero, and P. Garza, "EnBay: A Novel Pattern-Based Bayesian Classifier", IEEE Transactions On Knowledge And Data Engineering, Vol. 25(12), 2013.
22. PROMISE Software Engineering Repository, Available at: http://promise.site.uottawa.ca/ SERepository, accessed June 2014.
23. Paygude, Priyanka. (2020). Fault Aware Test Case Prioritization in Regression Testing using Genetic Algorithm. International Journal of Emerging Trends in Engineering Research. 8, 2112-2117. 10.30534/ijeter/2020/104852020.