INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue3)

Available online at: www.Ijariit.com

Software Defect Prediction Using Ensemble Learning Survey

Er. Ramandeep Kaur
Bahra Group of Institutes, Patiala
ramanpurewal04@gmail.com

Er. Harpreet Kaur
Bahra Group of Institutes, Patiala
preet.harry11@gmail.com

Er. Jaspreet Kaur
Bahra Group of Institutes, Patiala
jaspreetkaur.rb@gmail.com

Abstract- Machine learning is a science that explores the building and study of algorithms that can learn from the data. Machine learning process is the union of statistics and artificial intelligence and is closely related to computational statistics. Machine learning takes decisions based on the qualities of the studied data using statistics and adding more advanced artificial intelligence heuristics and algorithms

Keywords: Software default, machine learning, decision making, and promise dataset

I. INTRODUCTION

It is valuable to predict the software that is defect- prone. There have been many studies and learning approaches that are used to measure the performance of software. A meta-analysis of all relevant, high quality primary studies of defect prediction are used to determine that what factors influence predictive performance. The quality of the software can be measured with the different features such as cyclomatic complexity, design complexity, effort, time estimator, length of the program, operands, operators, line count etc. [12]

1.1 Software Engineering: Introduction

Software Engineering is defined as the systematic and well defined approach to the development, operation, maintenance and retirement of the software. By the word 'systematic' means that the methodologies used for the development of the software are repeatable. The goal of software engineering is to take software development closer to science and engineering that solves the problems of the clients and away from those approaches for development whose outcomes are not predictable. [12, 6]

1.1.1 Software Quality Attributes

Software Quality attributes can be defined as follows: [6]

1. **Functionality:** It is the capability that provides functions which meets the defined and implied needs of the software when it is used.
2. **Reliability**: It is the capability that maintains the specified level of performance.
3. **Usability**: The capability to be understood, learned and used.
4. **Efficiency:** It is the capability to measure the performance relative to the amount of resources used.
5. **Maintainability**: It is the capability to be updated and modified for purposes of making corrections, improvements or adaptation.
6. **Portability**: It is the capability to be adapted for different environments without applying actions.

1.2 Machine learning

Machine learning is a science that explores the building and study of algorithms that can learn from the data. Machine learning process is the union of statistics and artificial intelligence and is closely related to computational statistics. Machine learning takes decisions based on the qualities of the studied data using statistics and adding more advanced artificial intelligence heuristics and algorithms to achieve its goals. [5]

1.3 Data mining

Data mining is related with the discovery of new and interesting patterns from large data sets for analysis and executive decision making. Data mining is described as the union of historical and recent developments in statistics, artificial intelligence and machine learning. Data mining and machine learning are used together to study data and find previously-hidden trends or patterns within.[4,16,14,7]

1.3.1. Scope of Data Mining

- Automation in prediction of behavior and trends
- Automated discovery of previously unknown patterns

1.3.2 Data Mining Process

Data mining consists of five major elements:

- Extraction and transformation of data onto the data warehouse system.
- Run data on multidimensional database system in a managed way
- Providing data access to business analysts and other professionals
- Data analysing
- Presentation of data in useful and required formats such as tables and graphs.

LITERATURE REVIEW

A Okutan et al. [13] states different software metrics that are used for defect prediction and defines the set of metrics that are most important for predicting the defectiveness in the software module. The two more metrics i.e. number of developers and the source code quality are defined other than the promise data set. Experiments results that lines of code and lack of coding quality are the most effective metrics whereas coupling between objects and lack of cohesion of methods are less effective metrics on defect proneness.

Asir Antony Gnana Singh [2] discuss various dimensionality reduction methods (feature subset selection and feature ranking) To illustrate the applicability of feature subset evaluators (CFS, consistency, filtered) and feature rankers (Chi-squared, InfoGain) on large data sets.

Catal, U. Sevim, B. Diri [11] propose fully automated method for software module fault prediction. To implement the proposed work by incorporating X-means clustering (automatic generation of cluster number) and metrics threshold values (mean vector of each cluster is checked against metrics threshold).

C.Catal, U. Sevim, B. Diri [3] propose software fault prediction technique that uses clustering and metrics thresholds, to expel the accountability of human experts so as to make the purposed technique fully automated. To compare the results of the proposed technique with NB.

David Gray et al. [6] have explained the reason of significant pre-processing of data set for suitability of defect prediction. Researchers need to analyze the data that how it will be used by removal of constant attributes, repeated attributes, missing values and inconsistent instances. The experiments that have been used are based upon NASA metrics data program that results in errors findings and conclude that errors are mainly because of repeated data points.

Ferruh YIGIT [1] To analyse various text categorization methods to determine the individual features in the text. To propose a new feature selection method that is based on Info Gain and particle swarm optimization. To evaluate the performance of purposed method on the basis of classification accuracy, precision, recall and f-measure. To highlight the future work of feature selection method real world problems.

G Czibula et al. [4] focuses on the problem of importance during the time software evolution and maintenance for the problem of defect prediction. The quality of the software system is improved by identifying the defective software modules. Relational association rules are the extension of ordinal association rules that are being used in the prediction of software module that whether it is defective or not. It describes numerical range between the attributes that are commonly occur over a dataset by the experimental evaluation on the NASA datasets as well as a comparison is performed to similar existing approaches is provided.

Martin Shepperd et al. [15] have discussed about the factors having largest effect on the predictive performance of the software by conducting a Meta-analysis of all relevant and high quality primary studies of defect prediction on software module. The experimental results showed that the major factor is the researcher group instead of choice of classifier on the software performance.

Ming Li et al. [8] states that software quality can be controlled by software defect prediction. The defect prediction techniques used currently are based on large amount of historical data but in case of new projects and for many organizations historical data is often not available. In that case sample based methods for defect prediction can be used by selecting and testing a small percentage of module and after that build a defect prediction model to predict defect proneness of the other modules

.

M. Soudkhah , R. Janicki [20] analyze various algorithms by which weights are assigned to features and investigate those algorithms in which features have consistent weights.To illustrate pairwise comparisons method in order to rank the importance of features and show the results of accuracy improvement of classification.
To propose and test feature domain overlapping method of assigning weights to features.

Naeem Seliya,Taghi Khoshgoftaar[18] propose a constraint-based semi-supervised clustering approach that uses k-means algorithm.To aid the expert prediction and enhance the overall decision making.To show the results of semi supervised clustering are better than unsupervised.

Partha Sarathi Bishnu and Vandana Bhattacherjee [17] implement Quad tress based k-means clustering for software fault prediction.To show that clusters formed by above technique have maximum gain values.To evaluate the overall error rate and compare its performance with existing methods(NB,CS,CT,DA)

Qinbao Song et al. [19] describes the framework that comprises scheme evaluation and defect prediction components. Analysing the prediction performance for the given historical data sets is done by scheme evaluation and defect predictor constructs models according to the evaluated learning scheme and predicts defects of the software with new data according to the defined constructed model. It has been shown that different learning schemes should be used for different data sets.

Tim Menzies et al. [10] describes that how to improve the effort estimates of a project and defect predictions of a software module. The best thing can do to control cost and defects is to discard the needless functionality by making the lines of code to minimum. It has been seen that local treatments are always superior and different to the global treatments because data that appears to be useful in global context is often irrelevant to the local contexts.

CONCLUSION

That how to improve the effort estimates of a project and defect predictions of a software module. The best thing can do to control cost and defects is to discard the needless functionality by making the lines of code to minimum. It has been seen that local treatments are always superior and different to the global treatments because data that appears to be useful in global context is often irrelevant to the local contexts

REFERENCES

[1] A new feature selection method for text categorization based on information gain and particle swarm optimization. Yigit, Ferruh and Baykan, Orner Kaan. s.l. : IEEE,15. *Software Quality Analysis of Unlabeled Program.* Seliya, Naeem and Khoshgoftaar, Taghi M. March 2007, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 37, NO. 2.

[2] An Empirical Study on Dimensionality Reduction and Improvement of Classification Accuracy Using Feature Subset Selection and Ranking. Singh, D Asir Antony Gnana, Balamurugan, S Appavu Alias and Leaveline, E Jebamalar. s.l. : IEEE, 2012. International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET). pp. 102-108.

[3] Clustering and Metrics Thresholds Based Software Fault Prediction of Unlabeled Program Modules. Catal, Cagatay, Sevim, Ugur and Diri, Banu. s.l. : IEEE, 2009. Sixth International Conference on Information Technology: New Generations. pp. 199-204.

[4] Czibula, Gabriela, Zsuzsanna Marian, and IstvanGergelyCzibula. "Software defect prediction using relational association rule mining." *Information Sciences*264 (2014): 260-278.

[5] Dejaeger, Karel, Thomas Verbraken, and Bart Baesens. "Toward comprehensible software fault prediction models using bayesian network classifiers." *Software Engineering, IEEE Transactions on* 39.2 (2013): 237-257.

[6] Gray, David, et al. "The misuse of the nasa metrics data program data sets for automated software defect prediction." *Evaluation & Assessment in Software Engineering (EASE 2011), 15th Annual Conference on*.IET, 2011

[7] Hall, Tracy, et al. "A systematic literature review on fault prediction performance in software engineering." *Software Engineering, IEEE Transactions on* 38.6 (2012): 1276-1304.

[8] Li, M., Zhang, H., Wu, R., & Zhou, Z. H. "Sample-based software defect prediction with active and semi-supervised learning." *Automated Software Engineering* 19.2 (2012): 201-230.

[9] Li, Zhan, and Marek Reformat. "A practical method for the software fault-prediction." *Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on*. IEEE, 2007.

[10] Menzies, Tim, Butcher, A., Marcus, A., Zimmermann, T., &Cok, D. "Local vs. global models for effort estimation and defect prediction." *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*. IEEE Computer Society, 2011

[11] Metrics-driven software quality prediction without prior fault data. Catal, C, Sevim, U and Diri, B. 2010, Electronic Engineering and Computing Technology, Springer, pp. 189-199.

[12] Ohlsson, Niclas, Ming Zhao, and Mary Helander. "Application of multivariate analysis for software fault prediction." *Software Quality Journal* 7.1 (1998): 51-66.

[13] Okutan, Ahmet, and OlcayTanerYıldız. "Software defect prediction using Bayesian networks." *Empirical Software Engineering* 19.1 (2014): 154-181.

[14] Radjenović, Danijel, et al. "Software fault prediction metrics: A systematic literature review." *Information and Software Technology* 55.8 (2013): 1397-1418.

[15] Shepperd, Martin, David Bowes, and Tracy Hall. "Researcher bias: The use of machine learning in software defect prediction." *Software Engineering, IEEE Transactions on* 40.6 (2014): 603-616.

[16] Sherer, Susan A. "Software fault prediction." *Journal of Systems and Software* 29.2 (1995): 97-105.

.[17] Software Fault Prediction Using Quad Tree-Based K-Means Clustering Algorithm. Bishnu, Partha S and Bhattacherjee, Vandana. 2012, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE.

[18] *Software Quality Analysis of Unlabeled Program.* Seliya, Naeem and Khoshgoftaar, Taghi M. March 2007, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 37, NO. 2.

[19] Song, QinbaoJia, Z., Shepperd, M., Ying, S., & Liu, J. "A general software defect-proneness prediction framework." *Software Engineering, IEEE Transactions on* 37.3 (2011): 356-370.

[20] Weighted Features Classification with Pairwise Comparisons, Support Vector Machines and Feature Domain Overlapping. Soudkhah, M and Janicki, R. s.l. : IEEE, 2013. 22nd International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). pp. 172-177.

