Developing the improved software fault prediction approach using Feature Selection

Miss. S.V. Athawale, Dr.V.M.Thakare

Abstract- **Software fault prediction is a most valuable part in software quality to improve the performance and effectiveness of software.Classification is the method used in the process of software fault prediction. Fault prediction in software projects, improves thesoftware productivity. In this paper the method is proposed with a feature selection approach for software fault prediction. It extracts any feature which user wants to test for the fault. Feature is selected from N number of features to find whether there exists a fault or not. Feature selection is the process of identifying and removing irrelevant and redundant features from a dataset, so that only beneficial features are left for training the classification models.**

Keywords- **Software fault predictions, feature selection, feature ranking, instance reduction, data preprocessing, software metric selection.**

I. INTRODUCTION

Software fault prediction plays the important role in designing the different software fault prediction models for fault prediction. Software fault prediction models have been broadly classified into many models such as Kernel Regression Models, Linear Regression Models, Zero-inflated Poission Models. Each software fault prediction models have their own application with different behaviour to perform their task. Mostly used software fault prediction models are: Linear Regression Models (LR), Zero-inflated Negative Binomial models (ZINB), Negative Binomial Regression model (NBR). This paper, discusses five different software fault prediction approach such as Zero-inflated prediction model in software data [1], hybrid model reconstruction for cross-project defect prediction approach (HYDRA) [2], empirical studies of a two stage data pre-processing approach [3], learning-to-rank approach to software defect prediction [4] and fuzzy rulebased approach for software fault prediction [5]. These software fault prediction approach provide the better effectiveness, cross-project defect prediction, and prediction accuracy and quality. But these methods also have some problem so to overcome such problems improve version of software fault prediction models that is "Feature Selection" software fault prediction model is proposed here that depend upon the ranking and relevance analysis.

II. BACKGROUND

Many studies on software fault prediction models have been done to develop the software fault prediction approach in recent past years. Such approaches are: Zero-inflated prediction model in software fault data is proposed to develop a novel two-component approach to the restrictive existing models [1]. Hybrid model Reconstruction Approach (HYDRA) is designed for cross-project defect prediction which uses data from other projects to predict defects in a particular project, provides a new perspective to defect prediction [2]. Empirical studies of a two-stage data Developing the improved software fault prediction approach using Feature Selection pre-processing approach has been proposed that performs both feature selection and instance reduction in sequence, which combines feature selection and instance reduction to eliminate both irrelevant and redundant units in software datasets [3]. A learning-to-rank approach is presented to construct software defect prediction models by directly optimizing the ranking performance. The purpose of SDP for the ranking task is to predict which modules are likely to have most defects to allocate software quality enhancement efforts [4]. Fuzzy Rule- Based Approach for Software Fault Prediction has proposed that uses fuzzy rule-based classification systems. This proposed model presented a fuzzy rule-based framework for software fault prediction. Fuzzy modelling provides a very effective way to deal with vagueness/uncertainty that is associated with numerical measurements [5]. This paper introduces five software fault prediction i.e. Zero-inflated prediction model in software data, hybrid model reconstruction for cross-project defect prediction approach (HYDRA) , empirical studies of a two-stage data preprocessing approach, learning-to-rank approach to software defect prediction and fuzzy rule-based approach for software fault prediction. These are organized as follows. Section 1 Introduction. Section 2 discusses Background. Section 3 discusses previous work. Section 4 discusses existing methodologies. Section 5 discusses attributes and parameters and how these are affected on software fault prediction models. Section 6 proposed method and section 7 outcome result possible. Finally section 8 Conclude this review paper.

III. PREVIOUS WORK DONE

In research literature, many software fault prediction models have been studied to provide various fault prediction schemes and improve the performance in terms of cross-project fault prediction, effectiveness, data pre-processing. Roberta A.A. Fagundes, et al. (2016) [1] have worked on zero inflated prediction model in software-fault data that gives a more flexible structure to understand data. And provide the better performance and prediction accuracy. Xin Xia et al. (2016) [2] has proposed Hybrid model Reconstruction for cross-project defect prediction Approach (HYDRA) which addresses the all the challenges by iteratively learning new classifiers and compositions of classifiers to collectively better capture generalizable properties in every new iteration. Wangshu Liu et al. (2016) [3] has worked on empirical studies of a two-stage data pre-processing approach for Software Fault Prediction that is a novel two stage data pre-processing approach, which performs both feature selection and instance reduction in sequence, for SFP. Xiaoxing Yang et al. (2015) [4] has presents a learning-to-rank approach to construct software defect prediction models by directly optimizing the ranking performance. Pradeep Singh et al. (2015) [5] have proposed model presented a fuzzy rule-based framework for software fault prediction, Fuzzy modelling provides a very effective way to deal with vagueness/uncertainty that is associated with numerical measurements.

IV. EXISTING METHODOLOGIES

Many software fault prediction approach have been implemented over the last several decades. There are different methodologies that are implemented for different software fault prediction models i.e. zero-inflated prediction model in software-fault data, hybrid model reconstruction approach (hydra), empirical studies of a two-stage data pre-processing approach, learning-to-rank approach and fuzzy rule-based approach.

4.1 Zero-inflated prediction model in software-fault data: zero-inflated prediction model in software-fault data has presented a new model for software fault prediction. This proposed model developed a novel two-component approach as an alternative to the restrictive existing models. This approach combines one to classify whether or not a given element exhibits a zero response and one to predict a response value if that element is classified as exhibiting a non-zero response. This approach allows to combine parametric and nonparametric models to improve the prediction accuracy. This way provides a moreflexible structure to understand data [1].

4.2 Hybrid model Reconstruction Approach (HYDRA):

Hybrid model Reconstruction Approach (HYDRA) which addresses the all the challenges by iteratively learning new classifiers and compositions of classifiers to collectively better capture generalizable properties in every new iteration. This is the Massively Compositional Model for Cross- Project Defect Prediction efficient prediction models with two phases. HYbrid model Reconstruction Approach (HYDRA) for cross-project defect prediction, which includes two phases: genetic algorithm (GA) phase and ensemble learning (EL) phase. These two phases create a massive composition of classifiers. HYDRA contains two steps: model building step and prediction step. In the model building step, our goal is to build a cross project prediction model learned from instances in multiple source projects and the training target data. In the prediction step, we apply this model to predict if a new class/file/module in the target project has defects or not [2].

4.3 Empirical studies of a two-stage data pre-processing approach: Empirical studies of a two-stage data preprocessing approach for Software Fault Prediction that is a novel two stage data pre-processing approach, which performs both feature selection and instance reduction in sequence, for SFP. A novel two-stage data pre-processing approach which incorporates both feature selection and instance reduction. Specifically, in the feature selection stage, author first perform relevance analysis, and then propose a threshold-based clustering method, called novel threshold-based clustering algorithm, to conduct redundancy control. In the instance reduction stage, authorapplies random under-sampling to keep the balance between the faulty and non-faulty instances [3].

4.4 Learning-to-rank approach: Learning-to-rank approach is an efficient prediction approach with two aspects. The purpose of SDP for the ranking task is to predict which modules are likely to have most defects to allocate software quality enhancement efforts. The work includes two aspects: one is a novel application of the learning-to rank approach to real-world data sets for software defect prediction, and the other is a comprehensive evaluation and comparison of the learning-to-rank method against other algorithms that have been used for predicting the order of software modules according to the predicted number of defects. The LTR approach is implemented in Java. The LTR approach is applied to a wide range of real-world data sets, and given a comprehensive evaluation and comparison of LTR against other algorithms[4].

4.5 Fuzzy Rule-Based Approach: Fuzzy Rule-Based Approach for Software Fault Prediction, the proposed work uses fuzzy rule-based classification systems. This proposed model presented a fuzzy rule-based framework for software fault prediction. Fuzzy modelling provides a very effective way to deal with vagueness/uncertainty that is associated with numerical measurements. This method uses fuzzy rule-based classification systems. It has been observed that comprehension of rules becomes extremely difficult for human beings, if the rules involve a large number of atomic clauses. Lesser the number of antecedent clauses in a rule, the higher is its comprehensibility. Thus, feature selection is very important for fuzzy rule-based systems to make interpretable rules. Here, a fuzzy set theoretic framework is used because fuzzy reasoning is very similar to human reasoning and it is easy to comprehend. Fuzzy modeling provides a very effective way to deal with vagueness/uncertainty that is associated with numerical measurements. An advantage of fuzzy rule-based approach is that it can act as an interface between a numerical scale and a symbolic scale, which is commonly expressed by linguistic value [5].

V. ANALYSIS AND DISCUSSION

Zero-inflated prediction model combines one to classify whether or not a given element exhibits a zero response and one to predict a response value if that element is classified as exhibiting a non-zero response [1]. Hybrid model reconstruction for cross-project defect prediction approach (HYDRA) shows cross-project defect prediction which gives new perspective to defect prediction. HYDRA tunes a two layer hierarchical composition of a massive number of classifiers [2]. Empirical studies of a two-stage data preprocessing approach for Software Fault Prediction shows the elimination of both irrelevant and redundant units in software datasets. By the two-phase pre-processing, get a balanced, high quality dataset for training the classification models, which would improve the performance of fault prediction [3]. Learning-to-rank approach shows software defect prediction models by directly optimizing the ranking performance. The purpose of SDP for the ranking task is to predict which modules are likely to have most defects to allocate software quality enhancement efforts [4]. Fuzzy Rule-Based approach, shows a fuzzy rule-based framework for software fault prediction. Fuzzy modelling provides a very effective way to deal with vagueness/uncertainty [5].

Prediction models and approach	Advantages	Disadvantages
Zero-inflated prediction model	The classification and prediction models are constructed independently, according to each and a combination is established to provide the estimate and allowing different combinations for the classification and regression tasks.	It is not sensitive in the presence of situations of complex and nonlinear processes
Hybrid model reconstruction for cross project defect prediction approach (HYDRA)	ensures that projects can be planned, executed, controlled and improved. Reduce rework, standardize quality and continuously learn from past	HYDRA evaluate with datasets from less software projects. Some error that are unnoticeable may be present.
Two-stage data preprocessing approach	Improves the quality of software datasets used by classification models for software fault prediction and improves the performance of fault prediction.	The temporal issue requires consideration during data preprocessing. Based on experiments, either temporally adjacent software samples or highly coupled coding units may determine the usefulness of the software instances
Learning-to rank approach	This method provides easier way to capture the important factors related to defects by directly optimizing the ranking performance measure than by optimizing the prediction errors, especially when there are many coefficients to be optimized. LTR achieves better FPA results and LTR is more robust against noisy SDP data.	performed worse when the number of metrics is large.

Fuzzy rule based approach	This approach produce human understandable rules, which may be effectively used to guide software development Process. It achieves results competitive with other state of the art methods.	Prediction of cross company software defects is difficult because of the difference within company training data and cross company data.

TABLE 1: Comparisons between different software prediction approaches.

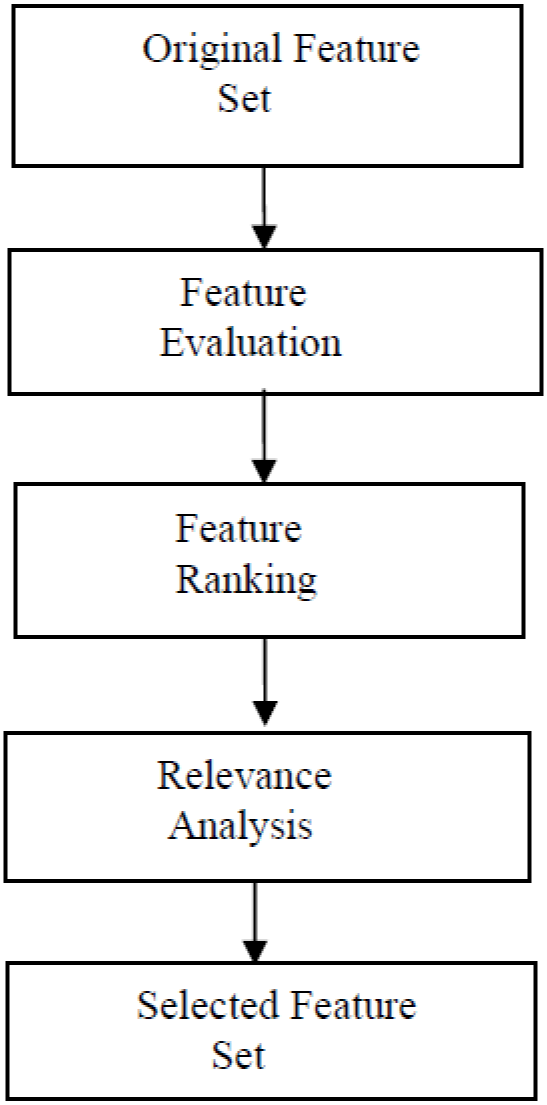
VI. PROPOSED METHODOLOGY

Software fault prediction approach is important and difficult task to analyse and discuss about various methods based on different parameters i.e. accuracy, quality, cost, time, flexibility, throughput, delay, capacity, effectiveness etc for different software fault prediction models. There are still problems which trouble in this field. New software fault prediction method called "Feature Selection" software fault prediction model for more effective and more accurate fault prediction model is propose here to overcome the problems of previous models. As this model is depend upon the high quality dataset for training the classification models, which would improve the performance of fault prediction. As Fault prediction in software projects improves the software productivity. One of the most important goals of such techniques is to accurately predict the modules where faults are likely to hide as early as possible in the development lifecycle. In feature selection, first perform relevance analysis to remove irrelevant features, and then propose a threshold based clustering algorithm for redundancy control to eliminate redundant features. Feature selection is the process of identifying and removing irrelevant and redundant features from a dataset, so that only beneficial features are left for training the classification models. A variety of feature selection methods have been developed, which can be roughly grouped into two categories: filter-based, and wrapper-based. Filter-based methods evaluate and select the most relevant features, based on the correlation between the features and class labels. Wrapper-based methods require feedback from the classification model, and compose the feature set iteratively, which may lead to high computational complexity.

Basic steps of algorithm:

Step1: The first step is performing relevance analysis to remove irrelevant features.

Step2: The second step is performing redundancy control to eliminate redundant features. This method extracts any feature which we want to test for the fault. Feature is selected from N number of features to find whether there is exist a fault or not. Diagrammatic representation of proposed method is shown as follows:



VII. OUTCOME AND POSSIBLE RESULT

In this way the proposed method is perform for the Feature selection software fault prediction. With the help of the feature ranking and threshold-based clustering in sequence the proposed method gives more effective and more accurate software fault prediction model and improve the performance of fault prediction.

VIII. CONCLUSION

This paper focused on the study of various software fault prediction approaches i.e. Zero-inflated prediction model in software data, hybrid model reconstruction for cross-project defect prediction approach (HYDRA), empirical studies of a two-stage data pre-processing approach, learning-to-rank approach to software defect prediction and fuzzy rule-based approach for software fault prediction. But there are some problems in accuracy and quality of software data so to improve this “Feature Selection” software fault prediction method is proposed here. Feature selection is the process of identifying and removing irrelevant and redundant features Original Feature Set Feature Evaluation Feature Ranking Relevance Analysis Selected Feature Set from a dataset, so that only beneficial features are left for training the classification models.

IX. FUTURE SCOPE

From observations of the proposed method the future work will include exact accuracy of software fault prediction with the help of effectiveness of metrics for SDP.

REFERENCES

1. Roberta A.A. Fagundes, Renata M.C.R. Souza , FranciscoJ.A. Cysneiros, “Zero-inflated prediction model in softwarefaultdata”, IET JOURNALS THE INSTITUTION OFENGINEERING AND TECHNOLOGY, VOL. 10, Iss. 1, JULY2016.

2. Xin Xia, David Lo, Sinno Jialin Pan, NachiappanNagappan, and Xinyu Wang,“ HYDRA: MassivelyCompositional Model for Cross-Project Defect Prediction”,IEEE TRANSACTIONS ON SOFTWARE ENGINEERING,VOL. 40, Iss. No. 10, October 2016.

3. Wangshu Liu, Shulong Liu, Qing Gu, Jiaqiang Chen,Xiang Chen and Daoxu Chen, “Empirical Studies of a Two-Stage Data Pre-processing Approach for Software FaultPrediction”, IEEE TRANSACTIONS ON RELIABILITY,VOL.65, Iss. No. 1, March 2016.

4. Xiaoxing Yang, Ke Tang and Xin Yao, “A Learning-to-Rank Approach to Software Defect Prediction”, IEEETRANSACTIONS ON RELIABILITY, VOL. 64, Iss. No. 1,March 2015.

5. Pradeep Singh, Nikhil R. Pal, Shrish Verma, and OmPrakash Vyas, “Fuzzy Rule-Based Approach for SoftwareFault Prediction”, IEEE TRANSACTIONS ON SYSTEMS,MAN, AND CYBERNETICS: SYSTEMS, VOL. 47, Iss. No. 5,MAY 2015