Notification No: 538/2023
Date of Award: 17/05/2023

Name of the Scholar: ROHIT VASHISHT
Name of the Supervisor: PROF. (DR.) SYED AFZAL MURTAZA RIZVI
Name of the Department/Centre: DEPARTMENT OF COMPUTER SCIENCE
(FACULTY OF NATURAL SCIENCES)
Topic of Research: HETEROGENEOUS CROSS PROJECT DEFECT PREDICTION

FINDINGS

The study includes all major parts of prediction of software defects using ML constantly evaluated and enhanced. The researcher particularly inclined to analyse and improve (1) the HCPDP model’s feasibility as compared with baseline WPDP model, (2) Problem of class imbalance in HCPDP model and its suitable corrections, (3) Extent of target coverage by HCPDP and lastly, (4) a novel SDP model fitted with optimal feature engineering procedures.

The first experiment determines if heterogeneous cross projects defect prediction is feasible, with the suggested method. HCPDP model was developed enabled with feature selection and ranking capacity along with five comparable ML classifiers for defect prediction. For model training, 10-fold cross validation was used.

When HCPDP is compared with WPDP, the significant findings, such as, XG Boosting outperforming all five classifiers and significant good performance of Kendall's Correlation (better than Spearman's Rho) ascertained that HCPDP model used in the research is aligning or can improved than WPDP base model.

SMOTE is used as the OS and RUS is used as the US to address CIL concerns in an asymmetrical training sample utilizing re-sampling approaches. Both SMOTE and RUS efficiently handle CIP on their respective sides as contrasted to the standard HCPDP. The experimental observations of the research have ensured that: firstly, SMOTE can produce better results than RUS since it avoids the over-fitting that occurs when using a simple OS method. Secondly, SMOTE efficiently manages CIP with a longer training time, but due to the elimination of some informative majority class instances when equalizing the observations of binary class with a shorter training duration, RUS produces unsatisfactory results in some scenarios.

The experiments to test the HCPDP model's validity revealed that the source project group SOFTLAB efficiently covered both the target project groups AEEEM and resynchronized to achieve the aim of 100% predicted coverage. With SVM, the greatest training accuracy of 0.97 is attained with AUC values of 0.892 and 0.889 were found in both pairs of the project under consideration. Hence, software metrics can develop DP models for the target application using defect records from other projects that differ from the target system (with or without defect information).

Here, an innovative approach is developed to solve the problem of partial defect prediction. Project pairs (ReLink, AEEEM) where partial failure prediction problems are found are overcome by continuously recording additional datasets from various SOFTLAB project groups until 100% DPC is achieved.

This study employs a 4-phase HCPDP modelling and a 3-phase WPDP to forecast defects in two or one project to compare performance efficiency with unsupervised learning.

Furthermore, this research demonstrates a logical analogy between heterogeneous and insidious project’s outcomes in terms of both learning approaches. Km++ clustering can be said to provide predictive output performance comparable to that of the LR method for both WPDP and HCPDP methods. The experiment to assess the performance level of WPDP to that of HCPDP showed HCPDP provides somewhat lower but equivalent performance to the traditional technique of DP, namely WPDP with application of unsupervised learning methods.

The importance of feature engineering uses a new 4-phase novel prediction model that adds deep FE techniques to extract the most relevant and powerful discriminated features to predict expected outcomes.

In addition, two new approaches, CBA and CBMMT, are being attempted to address the issue of class imbalances and evaluate the relationships between the characteristics of two multivariate programs.

For the WPDP and HCPDP, this study found that features cardinality was reduced by 55.61 percent and 48.35 percent, respectively. Experimental results show that the ability to compute the prediction performance with and without feature extraction is statistically significant compared to each DP in the program. The results suggest that extraction of features using deep learning methods has a larger effect on sampling prediction's accuracy rate than the data-driven FE method.

Based on the above comprehensive discussion on the experimental results of the thesis, the proposed machine learning based HCPDP model is enabled with following observations/findings:

- Comparable efficacy of defect prediction as the baseline WPDP model. Also, the HCPDP model's performance is comparable with baseline WPDP model in perspective of various classifiers. In this research, a GBM classifier performance was found to be the best.
- The proposed HCPDP model is effectively workable in defect prediction for projects with heterogeneous datasets.
- The proposed HCPDP model performs better with SMOTE (oversampling) to reduce CIP problems than RUS (under-sampling) and but, the former technique is taking longer training time.
- The proposed HCPDP-AE model applied with 4-phase deep learning based feature extraction approach gives reliably accurate result of defect prediction. The model can handle class imbalance and locating related features by using two novel methods, namely, CBA and CBMMT, respectively.
- From the feature engineering perspective, it is found that deep learning based feature extraction makes a more accurate defect prediction model.