e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:10/October-2021 Impact Factor- 6.752 www.irjmets.com

OVERVIEW OF SUPERVISED LEARNING TECHNIQUES FOR SOFTWARE DEFECT PREDICTION

N.Kalaivani*1, Dr.R.Beena*2

*1Associate Professor And Research Scholar, Department Of Computer Science, Kongunadu Arts And Science College, Coimbatore, Tamilnadu, India.

*2Associate Professor, Department Of Information Technology, Sri Ramakrishna College Of Arts And Science, Coimbatore, Tamilnadu, India.

ABSTRACT

Developing quality software has turn out to be a vital part of software development life cycle – hence building error -free software is essential. The main objective of software defect prediction is to identify whether a software module is error-prone or not. Error-prone software modules should be predicted before deploying the software. Thus, for building a prediction model, instead of considering all the attributes, it would be more beneficial to find out an appropriate set of attributes for predicting defects in software. Therefore, feature selection and feature extraction techniques seek to reduce the number of attributes in the dataset. The effectiveness of software prediction modules can be improved through reduced attributes set attained after feature selection and further used to identify defective modules in a specified set of inputs.

The main objective of choosing features in machine learning is to discover the finest set of features that permits to construct valuable models. Feature selection technique delivers a method of reducing number of attributes in the dataset, reducing calculation period and enlightening the performance of data prediction in machine learning applications. In this paper, we provide feature selection methods and dimensionality reduction methods are discussed.

Keywords: Defect Prediction, Supervised Learning, Dimensionality Reduction, Machine Learning, Heterogeneous Defect Prediction.

I. INTRODUCTION

A software should be developed in such a method that, it should satisfy all quality dimensions and produce error free software. A prediction model can be created by collecting historical information from a software project and forecasts faults in the similar project are known as With-in Project Defect Prediction [WPDP]. WPDP performed best, if there is enough quantity of historical data available to train models.

II. LITERATURE REVIEW

Jaechang Nam et al. [1] applied heterogeneous cross project defect prediction (HDP) to forecast faults across projects with heterogeneous metric sets. Authors applied feature selection techniques, which selects subsets of features by removing irrelevant and redundant features. Author applied matching metrics that have identical distribution of metrics with target metrics with source metrics.

Ramaswami et al. compared six filter feature selection algorithms and concluded that the results increase the predictive accuracy with low set of attributes [2]. Benchmarking of filter feature selection technique was subsequently carried out by applying various classification algorithms. The outcome of this study proves that there is increase in the accuracy of prediction with the existence of smallest number of features. The expected results illustrate a reduction in computational time and constructional cost in both training and classification phases of the performance model.

Kamal Bashir et al. [3] authors applied ranker feature selection [FS] techniques, Data Sampling [DS] and iterative – partition filter (IPF) to reduce high dimensionality, class imbalance and noisy features. Authors used classification techniques for predicting defect modules in software and compared its performance. From the results authors concluded that, Random Forest classification algorithm performs better when compared to other classification algorithms.

Ramani et al. employed pre-processing, data classification and classifier evaluation [4]. Finally, they conclude that extreme weighty features are selected by applying these techniques.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:10/October-2021 Impact Factor- 6.752 www.irjmets.com

Yu, Qiao, et al. developed an experimental work to discover the efficacy of feature selection for Cross Project Defect Prediction (CPDP), with feature subset selection and three feature ranking approaches [5]. Experiments were conducted using NASA and PROMISE datasets. The outcomes illustrate that, both feature subset selection and feature ranking approaches can increase the efficiency of CPDP.

Xia, Ye, et al. combined correlation analysis and ReliefF Feature Selection for selecting accurate features [6]. ANOVA (Analysis of Variance) analysis shows that, a new technique called ReliefF-LC (Linear Correlation Analysis) Feature Selection can increase the performance of defect prediction.

Kakkar, Misha, et al. for selecting features [7], authors applied Bat-based search algorithm and Random Forest algorithm for the prediction purpose.

Alsaeedi et al [8]. employed machine learning algorithms to predict software defects. Classification accuracy, f-measure and ROC-AUC metrics are used to evaluate performance measures of various machine learning algorithms. SMOTE resampling used to moderate the data inequality problem. Finally results proved that random forest, adaboost with random forest and bagging with decision tree algorithms performed well.

III. SOFTWARE DEFECT PREDICTION APPROACHES

Defect Prediction Approach	Concept	Advantage	Drawback
With-in Project Defect Prediction [WPDP]	A prediction model can be developed by considering past data from a software data repository and predict defects in the same software.	Prediction model works best when there is more enough past data in the repository to train models.	Prediction model is not suitable for other projects.
Cross Project Defect Prediction [CPDP]	Prediction model build for one project can be applied for other projects but both the projects must have similar number of metric sets.	Prediction model can be transformed from one project to other projects.	Prediction model not applicable for software projects which have different datasets.
Heterogeneous Cross Project Defect Prediction (HCPDP)	Increases the quality of software prediction model by applying prediction model of one project to other projects with imbalanced metric sets.	Prediction model works well for other projects with imbalanced datasets.	Nil

In order to achieve error-free software, it is important to predict error-prone modules prior to software deployment. To predict software defects efficiently, software development team should find out appropriate set of attributes which are relevant for defect prediction. To find out relevant set of attributes for constructing defect prediction model, feature selection, dimensionality reduction and feature extraction techniques are used.

IV. DIMENSIONALITY REDUCTION

Feature selection and dimensionality reduction methods seek to reduce the number of attributes in the dataset.

Difference between feature selection and dimensionality reduction

Feature Selection	Dimensionality Reduction
Chooses a set of features from the original space.	A transformation process is applied on features of original space and produce less features from original space.
Includes/excludes features present in the data without altering them.	Generates new mixtures of attributes.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:10/October-2021 Impact Factor- 6.752 www.irjmets.com

Methods for dimensionality reduction.

There are two methods to decrease the amount of features.

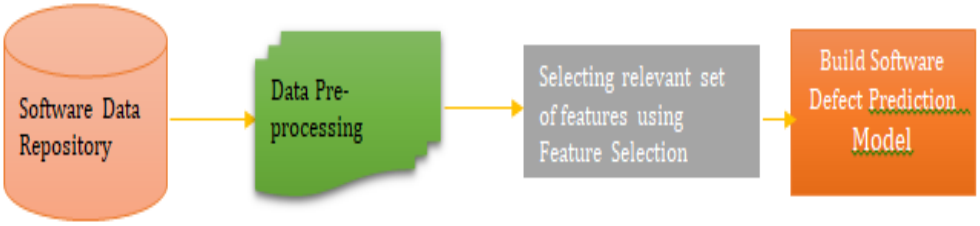
- Feature Extraction
- Feature Selection
- Feature Extraction

Produces new features based on the combinations of original features.

Approaches

- Principle Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Multidimensional Scaling.
- Feature Selection

A pre-processing step, which do not generate new features, but selects subset of features from the given set of features (i.e., original features) The main is to find the finest possible set of features for constructing a machine learning model. It can be employed in both supervised and unsupervised learning.

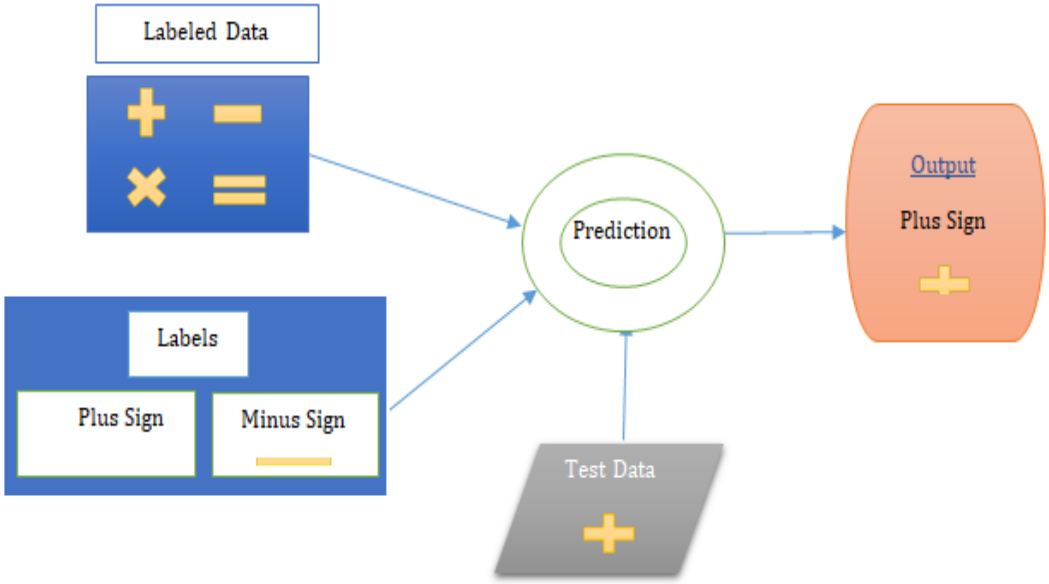


FEATURE SELECTION TECHNIQUES

a) Supervised Techniques
b) Unsupervised Techniques
c) Semi-supervised Techniques

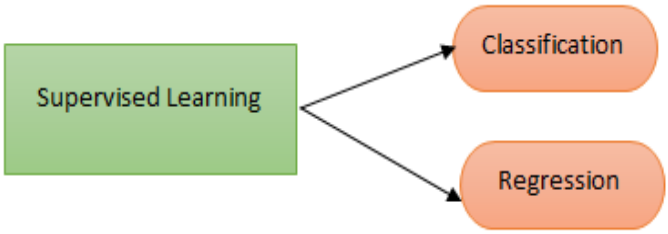
a) Feature Selection for Supervised Learning.

Supervised learning, also referred as supervised machine learning, is a subset of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms so as to categorize data or forecast results precisely. labeled data means some input data is previously marked with the accurate output. Machines given with the trained data, works accurately as a supervisor and predict the results correctly. As input data is feed into the model, it corrects its weights until the model has been fitted properly, which happens as part of the cross-validation process.



Types of supervised learning

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:10/October-2021 Impact Factor- 6.752 www.irjmets.com



b) Feature Selection for Unsupervised Learning

Unsupervised learning uses machine learning techniques to analyze and group unlabeled data according to their similarities. Machine learning algorithms determine unseen patterns in dataset without the necessity for human interference. Unsupervised machine learning algorithms [9] reduces the requirements of training labeled data.

c) Feature selection for Semi-Supervised learning

Training dataset is utilized with both labeled dataset and unlabeled dataset.

Difference between supervised learning and unsupervised learning techniques

Characteristics	Supervised Learning	Unsupervised Learning
Training Data	Trained with labeled dataset	Trained with unlabeled dataset
Response	Takes direct feedback to check if it is forecasting accurate result or not.	Does not provide any response for forecast of output.
Prediction	Output can be predicted.	Discovers unseen patterns in dataset.
Input	Input data given to the model along with the output.	Only input data given to the model.
Supervision	Requires supervision to train the model.	Do not require any supervision.
Types	Separated in to classification and regression.	Classified into clustering and association.
Outcome	Produces precise output.	Produces a lesser amount of precise output when compared with supervised learning.

Objectives of Feature Selection

- Removes unrelated and noisy features.
- Enables the machine lea
- Performing algorithm to train faster.
- Reduces the complexity of the model and makes it easier to interpret.
- Accuracy of the model gets increased if the right subset is chosen
- Increases the performance
- Reduces overfitting.

V. CONCLUSION

Developing a quality software system is a challenging factor in a software industry. On the other hand, the quality of the software system depends, that the software should be delivered on time, error-free and the system should meet all the requirements of an end-user. In order to produce an error-free software, software error should be identified and predicted at the earliest phase of the system life cycle itself. To predict the software faults different machine learning techniques can be applied. The main objective of this research study was to evaluate the previous research works with respect to software defect which applies supervised and unsupervised machine learning techniques for developing a software prediction model.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:10/October-2021 Impact Factor- 6.752 www.irjmets.com

VI. REFERENCES

[1] Nam, Jaechang, et al. "Heterogeneous defect prediction." IEEE Transactions on Software Engineering 44.9 (2017): 874-896.

[2] Ramaswami, M., and R. Bhaskaran. "A study on feature selection techniques in educational data mining." arXiv preprint arXiv:0912.3924 (2009).

[3] Bashir, Kamal, et al. "Enhancing software defect prediction using supervised-learning based framework." 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). IEEE, 2017.

[4] Ramani, R. Geetha, S. Vinodh Kumar, and Shomona Gracia Jacob. "Predicting fault-prone software modules using feature selection and classification through data mining algorithms." 2012 IEEE International Conference on Computational Intelligence and Computing Research. IEEE, 2012.

[5] Yu, Qiao, et al. "An empirical study on the effectiveness of feature selection for cross-project defect prediction." IEEE Access 7 (2019): 35710-35718.

[6] Xia, Ye, et al. "A new metrics selection method for software defect prediction." 2014 IEEE International Conference on Progress in Informatics and Computing. IEEE, 2014.

[7] Kakkar, Misha, et al. "Evaluating Missing Values for Software Defect Prediction." 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, 2019.

[8] Alsaeedi, Abdullah, and Mohammad Zubair Khan. "Software defect prediction using supervised machine learning and ensemble techniques: a comparative study." Journal of Software Engineering and Applications 12.5 (2019): 85-100.

[9] Li, Ning, Martin Shepperd, and Yuchen Guo. "A systematic review of unsupervised learning techniques for software defect prediction." Information and Software Technology 122 (2020): 106287.