Performance Analysis on Datasets and Heterogenous Defect Prediction through Machine Learning

Y Prasanth, B Yamini Supriya, M Tharun Kumar, P Y S Surya Teja

Abstract: ***Software defect prediction is an important factor which maintains high grade of software and reduces the cost of software development.In General, defect prediction identifies the modules that are defect-prone and then proceed for the testing phases.Literature survey is performed on Software defect prediction which is based on different machine learning techniques such as decision trees, neural network, Naive Bayes etc., This project presents the survey of various techniques to identify defects which also shows the accuracy between one defect to the other defect.To perform this experiment four NASA datasets have been used(defect data sets).These datasets are different in size and number of defective data. Finally, we end up with identifying which technique gives us more accuracy and less number of defects.***

Index Terms: software defect prediction, software reliability, fault prediction, defect data sets, accuracy.

I. INTRODUCTION

Beginning late, best software defect prediction approaches have been proposed and pulled in a ton.Software defects and deficiency in general affects the software reliability, software quality, product complexity.

It is very hard to achieve a fault-free or defect free software now-a-days ,even if the software is applied in cautious manner there are to be some hidden faults and mistakes.

In everyday language the terms Bug,Defect,Error,Failure are used inte changeably.Defect Prediction is an important factor in Software . Defect prediction is the most mature area in the field of software.To overcome Defects, many redundancy techniques such as Hard-ware redundancy, soft-ware redundancy, Information redundancy and Time redundancy can be used.Redundancy is performing the same functionality through the execution of different elements."Detection" is the key role in Software defect prediction ,detection is the observation of an error at a primary output, input simulation that creates an error as a result of fault. Several techniques are proposed to tackle(gear) the Soft-ware defect Prediction(SDP)problems.

Revised Manuscript Received on April 25, 2019.

Dr. Y Prasanth, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India.

B Yamini Supriya, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India.

M Tharun Kumar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India.

P Y S Surya Teja, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P, India.

The popularly known techniques are Machine Learning techniques(ML).The ML techniques are used mostly in SDP to predict the defected modules.Paper shows the study that talks about Naive Bayes(NB)classifier technique,Multilayer perceptron, Support Vector Machine(SVM)technique, Decision Tree classifier(DT).These ML classifiers are been applied to the dataset works and in computation the paper compares between one algorithm to other . Comparison shows the Error Rate and Accuracy.The outline of these ML algorithms is being shown and also we showcased a dataset and assessment approach.Experimentation Results are shown in conclusions.

II. LITERATURE REVIEW

Watanabe proposed the estimation pay approach for Cross project defect prediction. The metric pay changes an objective dataset like a source dataset by utilising the conventional estimation respects. To assess the execution metric pays, Watanabe et al. gathered 2 deformation datasets with an equivalent estimation set from two programming assignments and a brief span later organized CPDP.

Rahman et al. assessed the CPDP execution to the degree cost effectiveness and concerned that the longing execution of CPDP resembles WPDP .For the unequivocal examination,He gathered nine data-sets with similar strategy met-ric sets. Fukushima composed a distinct examination of just-in-time defect-prediction in the 'CPDP' settings.They utilized 16 datasets with a practically identical estimation set. The 11 datasets were given by Kamei in any case, 5 undertakings were starting late collected with a similar estimation set of the 11 datasets. Regardless, gathering datasets with a cost estimation set may oblige CPDP. For instance if existing bending datasets contain object-composed estimations, for example, CK estimations, amassing a practically identical thing masterminded estimations is unimaginable for activities . Hans J. Lenz proposed an algorithm which indicates No.of Clusters based on similarity matrix.This idea is implemented form spectral clustering,clustering, process on stochastic process on graph and Cramer theory.

Yvan Saeys introduced ensemble methods for feature selection.

He showed constucting ensemble feature selection techniques of feature ranking and feature subset selection.

Nam et al. adjusted a front line exchange learning system known Transfer Component Analysis (TCA) and also proposed TCA+. Usage of 8 datasets in two social events, ReLink and AEEEM are been used

Regardless, Nam et al. couldn't prompt CPDP among ReLink and AEEEM in light of the manner in which that they have heterogeneous estimation sets. Since errand pool an identical estimation set is compelled, planning CPDP utilizing an undertaking bunch with an equivalent estimation can be kept good.In that limit, we can't really lead CPDP for of the bending data-sets by utilizing staying (82%) data-sets in the PROMISE store. CPDP assesses driven by Canfora et al. moreover, Pani-chella et al. utilized 10 Java undertakings just with a tantamount estimation set from the PROMISE vault Zhang et al. proposed the no19 matter how you look at it appear for CPDP . Regardless, the extensive imperfection want model might be difficult to apply for undertakings with heterogeneous estimation sets since the far reaching model uses 26 estimations including code estimations, object-orchestrated estimations, and strategy estimations. Everything considered, the model must be reasonable for target datasets with an equivalent 26 estimations. For the situation where the objective experience has not been made in article engineered tongues, a general model made utilizing object-organized estimations can't be utilized for the objective dataset.

He et al. kept an eye out for the hindrances because of heterogeneous estimation sets in CPDP examines recorded above.Their procedure, CPDP-IFS, utilized stream trademark vectors of an occasion as estimations. The guess execution of their best methodology is in every practical sense undefined to or solid in improving ordinary CPDP models.. In any case, the rationality by He et al. isn't separated and WPDP.

Regardless of the way the best rationality is useful to increase the standard CPDP models, the examination might be delicate because the longing execution of a common CPDP is consistently unfathomably low. Additionally, He et al. facilitated examinations on just 11 extends in 3 dataset social events

Normal E.Fenton proposed in illustration of these points the “Goldilock’s Conjecture”.Issues that surrounds the “Goldilock’s Conjec-ture” illustrates how difficult is the defect prediction is and how easy to commit modeling mistakes.

Naeem Seliya says feature selection is important in defect prediction also implies that selecting the correct sets of software metrics for defect prediction is important. Working with a smaller set of metrics for software quality modeling is more attractive than working with a large number of metrics.

III. USED MACHINE LEARNING TECHNIQUES

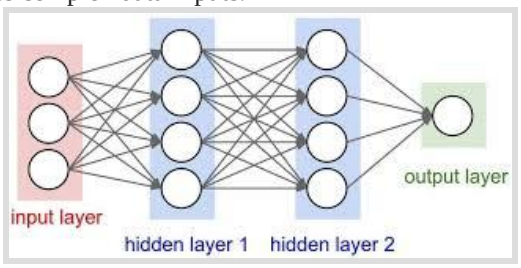
The study shows the assessment of Machine learning techniques namely, Naive bayes(NB),Multi layer perceptron, Support Vector Machine,Decision Tree(DT). This paper shows the performance accuracy,Error Rate and capability of ML algorithms .Also Shows us the comparative analysis of the algorithms. These algorithms show the difference between Accuracy and Error Rate.Here is the detailed description of the algorithms used:

Naive Bayes (NB):

Naive Bayes is a effective and a probabilistic classifier based on Bayes theorem Also gives an illusion b/w the features.

Artificial neural networks(ANN):

The neural network itself is not an ,algorithm, but rather it is a frame-work for many different ML algorithms to work and process complex data inputs.



ANN implements the signal at a connection between artificial neuron is the real number, and the output of each neutron is computed by some non-linear function of the sum of i/p’s

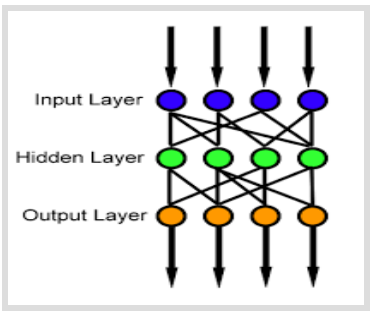
Multilayer perceptron(MLP):The Multi-layer perceptron differs from the simple perceptron in many ways.An MLP is a class of “feed-forward neural network”.Consist of three layers:

A)The input layer

B)The hidden layer and

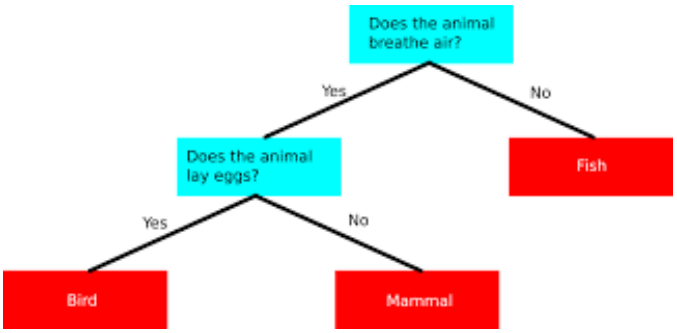
C)The output layer

A supervised learning technique namely “back propagation” is been utilised byMutilayer perceptron for training.



Decision Tree(DT):

A decision tree is a tree like structure which uses a tree-graph or model of decisions and their possible consequences.Decision tree uses a branching method to illustrate output of decisions.In decision ,tree, each branch represents a possible decision or reaction.

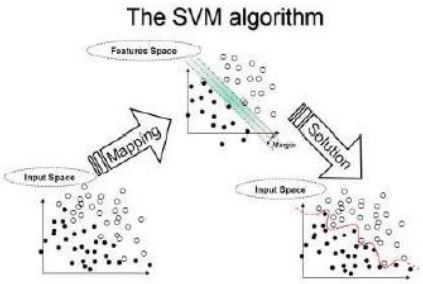


The result of a DT will be called Decision -nodes and Leaf-nodes.A Decision node will be having two or more branches.

Support-vector machine(SVM) :

SVM is a ML technique to divide data which tries to reduce the gap between the categories.

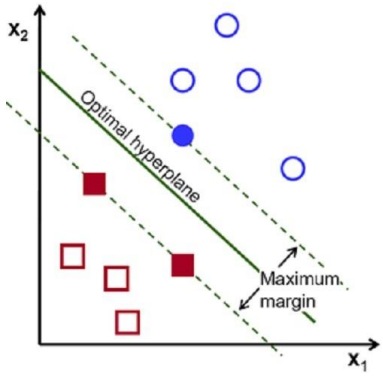
SVM is supervised algorithm which is assocated with learning algorithms that analyse data used for classification&regression. SVM model is the representation of examples as points in space.



Maximum Margin Classifier

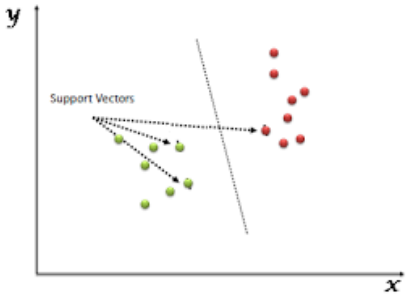
A margin classifier ,gives an associated distance from the decision boundary.

Maximum Margin Classifier divides data, if the data is linearly separable. But due to some reasons it cannot be applied to the data where no clear linear boundary is possible.



Support Vector Classifier

SVC is an supervised machine Learning algorithm.Can be used for classification and regression.



IV. RESULTS

The proposed system is mostly implemented by using the four classification algorithms. Among the classification "MultiLayer perceptron" is most widely used and performance is improved based on the accuracy and error rate.

A)**Confusion-Matrix**

It is a table that is designed to record the performance of ML algorithm.Every row of the confusion matrix reperesents the actual class, column indicates the instance in a predicted class or conversely.this gives the result as true positive, true negative, false positive, false negatives

B)**Accuracy**

Accuracy is the state of being correct.Results will. Be both True positive,True negative(TN,TF) among the overall No.of examined instances. 1 and 0 are the best and worst accuracy values respectively. Formula:

A=(TP+TN)/(FP+FN+TP+TN)

C)**Precision**

Precision is the state of being exact.

Predictive positive value

Precision is be calculated by :

Number of correct predictions/Total number of positive predictions. 1 and 0 are the best and worst ever precisions and can be calculated as:

Pre= TP/(FP+TP)

D) **Recall**

TP rate

Re-call is calculated by:

Number of +VE (Positive predictions) /the total number of positives.

Computed by

R= TP/(FN+TP)

E) **F-Measure**

F-Measure is the weighted mean of recall and precision.

F-Measure combines both precision measures and recall in one measure to compare ML algorithms

calculated by the following Formula:

F-Measure= (2*precision*recall)/(precision+recall)

F. Root-Mean-Square-Error (RMSE)

Root Mean square error is the measure for computing the efficiency (or) Performance of the prediction model. Measure the difference b/w the actual and predicted values.if the predicted value is X and the actual value is XP then RMSE is as
shown below:

RMSE = SQRT(1/N)*SIGMA n to i=1 (xi-xpi)^2

V. DATA SETS

	Navie Bayes	Multilayer Preceptron	SVM	Decision Tree
Accuracy	0.814135	0.81966	0.817471	0.81674
Error rate	0.18586	0.180339	0.182526	0,18325

VI. EXPERIMENT RESULT

This CASE STUDY used WEKA which is a good machine learning tool which evaluatesML algorithms in software Defect prediction problem. A cross validation of 10 folds is been used for each dataset. The accuracy and the Error Rate of these algorithms are shown. The three ML algorithms achieved an accuracy rate and error rate. Among the classification ,SVM is most widely used and performance is improved based on the accuracy and error rate.

VII. CONCLUSION

Defect prediction is most widely implemented in many places and also in heterogeneous systems. Datasets are used such as bank dataset and supermarket datasets are utilised and performance is show in this paper.

REFERENCES

1. G. Czibula, Z. Marian,and I. G. Czibula, “Software defect prediction using relational association rule mining,” Information Sciences, vol. 264, pp. 260–278, 2014.
2. X. Y. Jing, S. Ying, Z. W. Zhang, S. S. Wu, and J. Liu, “Dictionary learning based software defect prediction,” In Proceedings of the 36th International Conference on Software Engineering, 2014, pp. 414–423.
3. T. Menzies, J. Greenwald, and A. Frank, “Data mining static code attributes to learn defect predictors,” Software Engineering, IEEE Transactions on, vol. 33, no.1, pp. 2–13, 2007.
4. I. H. Laradji, M. Alshayeb, and L. Ghouti. Software defect prediction using ensemble learning on selected features. Information and Software Technology, 58:388–402, 2015.
5. Y. Ma, G. C. Luo, X. Zeng and A. Chen, “Transfer learning for cross company software defect prediction,” Information and Software Technology, vol. 54, no. 3, pp. 248-256, 2012.
6. J. Nam, S. J. Pan and S. Kim, “Transfer defect learning,” Proceedings of the 35th International Conference on Software Engineering . San Francisco, 2013, pp. 382-391.
7. J. Nam and S. Kim, “Heterogeneous defect prediction,” In Proceedings of the 10th Joint Meeting on Foundations of Software Engineering, 2015, pp. 508–519.
8. X. Y. Jing, F. Wu, X. Dong, F. Qi, and B. Xu, “Heterogeneous cross-company defect prediction by unified metric representation and cca-based transfer learning,” In Proceedings of the 10th Joint Meeting on Foundations of Software Engineering, 2015, pp. 496–507.
9. Y. Jiang, M. Li, and Z.-H.Zhou, “Software defect detection with rocus,” Journal of Computer Science and Technology, vol. 26, no. 2, pp. 328– 342, 2011.
10. D. Ryu, O. Choi, and J. Baik, “Value-cognitive boosting with a support vector machine for cross-project defect prediction,” Empirical Software Engineering, vol. 21, no. 1, pp. 43-71, 2016.
11. S. Watanabe, H. Kaiya, and K. Kaijiri. Adapting afault prediction model to allow inter languagereuse. In Proceedings of the 4th International Workshop on Predictor Models in Software Engineering, pages 19{24, New York, NY, USA, 2008.ACM.
12. F. Rahman, D. Posnett, and P. Devanbu. Recalling the \imprecision" of cross-project defect prediction. In Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering, New York, NY, USA, 2012.ACM.
13. T. Fukushima, Y. Kamei, S. McIntosh, K. Yamashita, and N. Ubayashi.An empirical study of just-in-time defect prediction using cross-project models. In Proceedings of the 11th Working Conference on Mining Software Repositories, pages 172{181, New York, NY, USA, 2014.ACM.
14. Y. Kamei, E. Shihab, B. Adams, A. Hassan, A. Mockus, A. Sinha, and N. Ubayashi. A large-scale empirical study of just-in-time quality assurance.Software Engineering, IEEE Transactions on, 39(6):757{773, June 2013.
15. B. Turhan, T. Menzies, A. B. Bener, and J. Di Stefano.On the relative value of cross-company and within-company data for defect prediction.Empirical Softw. Eng., 14:540{578, October 2009.
16. Y. Ma, G. Luo, X. Zeng, and A. Chen.Transfer learning for cross-company software defect prediction.Inf. Softw. Technol., 54(3):248{256, Mar. 2012.
17. J. Nam, S. J. Pan, and S. Kim.Transfer defect learning. In Proceedings of the 2013 International Conference on Software Engineering, pages 382{391, Piscataway, NJ, USA, 2013. IEEE Press.
18. G. Canfora, A. De Lucia, M. Di Penta, R. Oliveto, A. Panichella, and S. Panichella.Multi-objective cross-project defect prediction. In Software Testing, Verification and Validation, 2013 IEEE Sixth International Conference on, March 2013.
19. T. Menzies, B. Caglayan, Z. He, E. Kocaguneli, J. Krall, F. Peters, and B. Turhan. The promise repository of empirical software engineering data, June2012.

20. A. Panichella, R. Oliveto, and A. De Lucia.Cross-project defect prediction models: L'union fait la force. In Software Maintenance, Reengineering and Reverse Engineering (CSMR-WCRE), 2014 Software Evolution Week - IEEE Conference on, pages 164{173, Feb 2014.
21. F. Zhang, A. Mockus, I. Keivanloo, and Y. Zou.Towards building a universal defect prediction model.In Proceedings of the 11th Working Conference on Mining Software Repositories, MSR 2014, pages182{191, New York, NY, USA, 2014.ACM.
22. P. He, B. Li, and Y. Ma.Towards cross-project defect prediction with imbalanced feature sets.CoRR, abs/1411.4228, 2014.
23. I. Guyon and A. Elissee_. An introduction to variable and feature selection. J. Mach. Learn. Res., 3:1157{1182, Mar. 2003

AUTHORS PROFILE

Prasanth Yalla received his B.Tech Degree from Acharya Nagarjuna University, Guntur (Dist), India in 2001, M.Tech degree in Computer Science and Engineering from Acharya Nagarjuna University in 2004, and received his Ph.D. degree in CSE titled “A Generic Framework to identify and execute functional test cases for services based on Web Service Description Language” from Acharya Nagarjuna University, Guntur (Dist), India in April 2013.

He was an associate professor, with Department of Information Science and Technology in KL University, from 2004 to 2010. Later he worked as Associate professor, with the department of Freshman Engineering from 2011 in KL University. Presently he is working as Professor in the department of Computer Science & Engineering in KL University. Till now he has published 9 papers in various international journals and 4 papers in conferences. His research interests include Software Engineering, Web services and SOA. He taught several subjects like Multimedia technologies, Distributed Systems, Advanced Software Engineering, Object Oriented Analysis and design, C programming, Object-Oriented programming with C++, Operating Systems , Database management systems, UML etc. He is the Life member of CSI and received “Active Participation- Young Member” Award on 13-12-13 from CSI.