Improving Software Development Process through Data Mining Techniques of Unsupervised Algorithms

Dr. R. Durga
Assistant Professor,
Department of Computer Science,(VISTAS),
Chennai, India.

Ch. Kishore Kumar
Research Scholar,
Department of Computer Science, (VISTAS),
Chennai, India.

***Abstract*:- The important aim of software program improvement is to supply first-rate software program effectively and in the least quantity of time whenever possible. To reap the previous goal, builders frequently favor to reuse current frameworks or libraries as an alternative of growing comparable code artifacts from scratch. The difficult thing for builders in reusing the current frameworks or libraries is to recognize the utilization patterns and ordering policies amongst Application Programming Interfaces (APIs) uncovered through these frameworks or libraries, due to the fact many of the existing frameworks or libraries are no longer properly documented. Incorrect utilization of APIs can also lead to violated API specifications, main to protection and robustness defects in the software. Furthermore, utilization patterns and specs would possibly exchange with library refactoring, requiring modifications in the software that reuse the library. Data mining methods are utilized in constructing software program fault prediction fashions for enhancing the software program quality. Early identification of high-risk modules can aid in great enhancement efforts to modules that are possibly to have a excessive quantity of faults. This paper affords the data mining algorithms and strategies most generally used to produce patterns and extract fascinating data from software program engineering data. The strategies are equipped in seven sections: classification trees, affiliation discovery, clustering, synthetic neural networks, optimized set reduction, Bayesian faith networks, and visible records mining can be used to acquire excessive software program reliability.**

Keyword: Software Reliability, Testing, Defect prediction, data mining, Machine learning techniques.

I. INTRODUCTION

A standard software program improvement method, the work is cut up into distinct levels with unique activities in every, with the aim of improving making plans and management. The maximum normally used methodologies include waterfall, prototyping, iterative and incremental improvement, spiral improvement, fast utility improvement, excessive programming and numerous forms of agile method. While a life-cycle version is a greater trendy time period for class of methodologies, a software program improvement "method" is frequently synonymous to a selected method selected through a selected organization. A variety of such frameworks have developed over the years, every with its personal recognized strengths and weaknesses. One software program improvement method framework isn’t always appropriate to be used through all projects. Each of the to be had method frameworks are high-satisfactory acceptable to precise varieties of projects, primarily based totally on numerous technical, organizational, venture and group considerations. Such contrasting improvement paradigms and the difficult dependencies that they invent growth the complexity of software program systems. This slows down improvement and maintenance, reasons faults and defects and sooner or later ends in a growth in price of the software program. Organizations frequently fail to recognize how their method influences the nice of the software program that they produce. This is particularly due to the issue innate in discovery and dimension. Although software program metrics have lengthy been the de-facto trendy for the evaluation of software program excellent and improvement processes, their drawbacks are numerous. The overreliance on metrics that may be without problems received and understood, utilization of metrics that seem thrilling however stay inappropriate and uninformative and the issue in acquiring honestly treasured metrics are however to call a few. Data mining is described because the method of coming across formerly unknown and doubtlessly beneficial records from facts collections. Thus utilizing facts mining in software program trying out with the purpose of software program development has piqued the hobby of researchers worldwide. There are numerous demanding situations that emerge in mining software program repositories. The essential ones being, managing the inherent complexity and sheer quantity of the software program engineering facts. Data mining concentrates on running with massive portions of facts to offer a pattern. In phrases of purchaser facts it's far very beneficial to result in achieve successful marketing. So on occasion it's far violating the records safety regulation through proving unknown relationships in facts.

II. LITERATURE SURVEY

1. Research Progress on Software Engineering Data Mining Technology:

At present, with the dimensions enlargement of computer software program, best rely upon guide for software program improvement, protection and different models is extra difficult. Data mining era can boost up the rate of software program improvement, and might in many databases locate precious facts. Makes in-intensity research on software program engineering facts mining era, and introduces the affect of facts mining era. Software engineering facts mining era is to apply current era or new facts mining set of rules in huge databases, and is the method of gathering precious records for software program builders via a chain of steps, together with selection, analysis, formulation. It is a method of clean grasp and control of software program improvement. Software builders need to accumulate the required facts, that is the exercise of software program improvement industry. To complete the paintings, they extracted the specified facts records from huge quantities of facts, and the method of gathering and selecting. records is the method of facts mining. At present, facts mining era has been extensively used in software program testing. Data Mining Techniques used in Software Engineering A Survey Standard software program improvement system has numerous stages; every with its own importance and dependency at the different. Each degree is frequently complicated and generates a huge form of information. Using information mining strategies, they can discover hidden styles from this information, degree the effect of every degree at the different and collect beneficial statistics to enhance the software program improvement system. The insights won from the extracted information styles can help software program engineers to predict, plan and realize the numerous intricacies of the project, permitting them to optimize destiny software program improvement activities. As each degree with inside the improvement system involves a positive final results or goal, it will become critical to choose the first-class information mining strategies to obtain these desires efficiently. In , they surveyed the to be had information mining strategies and proposed the maximum suitable strategies for every degree of the improvement system. They additionally speak how information mining improves the software program improvement system in phrases of time, cost, resources, reliability and maintainability.

III. EXISTING SYSTEM

Software defect prediction (SDP), which classifies software modules into defect prone and not-defect-prone categories, provides an effective way to maintain high quality software systems. Most existing SDP models attempt to attain lower classification error rates other than lower misclassification costs. However, in many real-world applications, misclassifying defect-prone modules as not defect-prone ones usually lead to higher costs than misclassifying not-defect-prone modules as defect-prone ones. The data mining techniques, the study didn’t provide a better accuracy for software defect prediction. The study uses only three methods for extracting the feature from the large data sets. These are not enough for better prediction need to explain more methods or algorithm in both feature extraction and classification methods.

IV. PROPOSED SYSTEM:

Software Defects Prediction Using E system:

Data mining Techniques used to find the defects that are present in the software product during testing of each phases. Different statistical methods or algorithms are used in feature extraction phase to improve the accuracy of the defect prediction. Software Defect Prediction is an important aspect in order to ensure software quality. The evolutionary aid vector device (ESVM) is an more or optimized shape of not unusual help vector device approach and it represents optimized algorithms for training to lean several elegance in addition to regression policies from datasets below interest. As for example, the Evolutionary assist Vector machine (ESVM) may be probably hired for learning various classifier techniques which encompass polynomial classifier; radial basis feature (RBF) based totally definitely classifiers and multi-layer perceptions (MLP) styles of classifiers. Inception ally, the evolutionary SVM (ESVMs) have been first endorsed with the aid of the usage of way of (Vapnik, 1960) for displaying facts type and private presently become a location of strenuous take a look at our inside the purple for upgrades within the strategies and hypothesis together with conservatories to expose off regression and estimation of density. Intrusion Detection the use of Proposed ESVM Mining Module everyday and assault internet net web page site visitors are categorized with the useful resource of the ESVM, in the direction of schooling ESVM learns the everyday and assault styles from the training record. In attempting out ESVM differentiate the attack and ordinary internet web site online website site visitors the usage of discovered out styles. schooling of ESVM on this studies paintings a drastically robust Evolutionary beneficial resource Vector device (ESVM) mechanism has been hired for training of mining module. This training technique classifies a given statistics difficulty x € Rn via manner of assigning a label y € normal, ICMP, TCP, UDP, Smurf, Port test, Land, HTTP, session, IP. Wherein x is the set of inputs to the manual Vector gadget, Rn is relation among n attributes, y is the output produced through the EMCSVM which consist frequently different types of classes which encompass one regular elegance and nine assault instructions. Education report for Evolutionary help Vector tool (ESVM) is made from schooling statistics set. Schooling information set file carries 20 attributes and 10 styles of schooling. ESVM educated using the education statistics set and weight values then produce the version record, this version record Is used to classify the handiest-of-a-type styles of assaults, Proposed system use five open source datasets from NASA Promise Data Repository to perform this comparative study. For evaluation, three widely used metrics: Accuracy, F1 scores and Areas under Receiver Operating Characteristic curve are used. It is found that Artificial Neural Network outperformed all the other dimensionality reduction techniques.

The dataset, used for Software Defect prediction in the project, is taken from NASA Promise Repository. All the 5 data sets have 22 attributes, though each having a different number of instances. Decision Tree classifier is used to make the model learn from the test set and then the model is tested on the training set and the performance measures are calculated. However, having so many attributes and instances can lead the model to overfit. Hence, we first reduced the dimensionality of the data to a set of 8 cumulated features using 4 different techniques and then trained the model using Decision Tree classifier. A detailed comparison was then made based on the performance metrics that include Accuracy, F1-Scores and Area Under the Receiver Operating Characteristics (ROC).

VI. PERFORMANCE MEASURES

Sno. Type Of Bugs	Predicted Buggy	Predicted Clean
TRUE BUGGY	TP	FN
TRUE CLEAN	FP	TN

A. Accuracy This refers to the ratio of correctly predicted instances of the test set to the total number of instances of the test set.

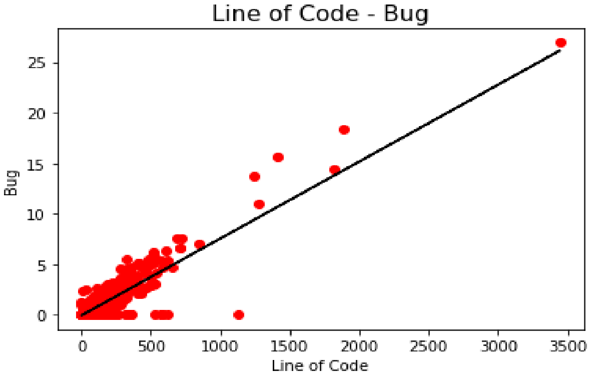
Accuracy = (TP + TN) / (TP + FN + FP + TN)

B. F1 scores At times, accuracy paradox can lead to misinterpretation of the results, hence we take another performance metrics called F1 score into consideration. F1 score is the harmonic mean of Precision and Recall, which are also calculated from the confusion matrix. Precision is the ratio of actual correctly predicted positive (buggy) instances to the total number of predicted positive instances (Precision =). TP TP + FP Recall is also known as Sensitivity.

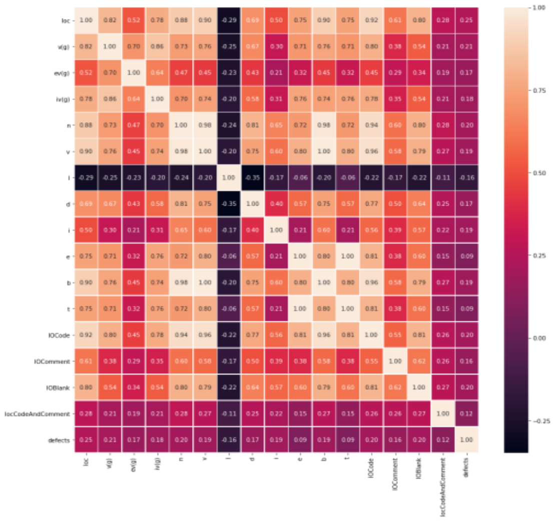
VII.VALIDATION METHOD

Here used hold out cross validation method to validate the data set. Since all the data sets used had quite a large number of instances, the training set and test set were divided in the ratio 3:1. The training set was used to train the classifier and then the model was validated on the test set.

VIII. RESULTS



Area Under the Curve (AUC) The performance of the predicted models was evaluated by plotting the Receiver Operating Characteristics (ROC) curve and evaluating the area under the curve. ROC curve, which is defined as a plot of sensitivity on the y-coordinate versus its 1-specificity (it is defined as the ratio of predicted non faulty classes to the number of classes actually non faulty) on the x coordinate, is an effective method of evaluating the quality or performance of predicted models.



. F1 scores At times, accuracy paradox can lead to misinterpretation of the results, hence we take another performance metrics called F1 score into consideration. F1 score is the harmonic mean of Precision and Recall, which are also calculated from the confusion matrix. Precision is the ratio of actual correctly predicted positive (buggy) instances to the total number of predicted positive instances (Precision =). TP TP + FP Recall is also known as Sensitivity.

Recall is the ratio of actual correctly predicted positive (buggy) instances to the total number of actual positive instances (Recall =) TP TP + FN Taking the harmonic mean, we get F1 score = Recall + Precision 2*Recall*Precision

IX. CONCLUSION

The current generation data mining and machine learning technology is widely used in developing new software's and testing. data mining in testing can get better the preservation competence of software system, , increases system stability. Machine learning techniques combined with testing techniques helps to produce high quality software. This paper discusses various reliability estimation techniques and software defect prediction using machine learning techniques.

X. REFERENCES

[1] Emad Alsukhni et al, "A New Data Mining-Based Framework to Test Case Prioritization Using Software Defect Prediction", International Journal of Open Source Software and Processes 8(1):21-41 • January 2017.

[2] Yanguang Shen, et al., "Research on the Application of Data Mining in Software Testing and Defects Analysis", 2009 Second International Conference on

[3] A.Kanagaraj, et.al, "A Study on Technologies Used in Ubiquitous and Pervasive Computing", International Journal of Advanced Research in Computer Science, Volume 3, No.3, May-June 2012.

[4] Last M.,et.al "Data Mining for Software Testing", In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA, 2005.

[5] Nadhem et.al, "The use of data Mining Techniques for Improving Software Reliability", International Journal of Advanced

Research in Computer Science, Volume 4, No. 4, March-April 2013.

[6] Halim. et.al,., "Similarity distance measure and prioritization algorithm for test case prioritization in software product line testing", Journal of Information and Communication Technology, 18(1), 57-75, 2018.

[7] Mark Last, et.al, "Using Data Mining For Automated Design of Software Tests", Department of Information Systems Engineering, Ben-Gurion University of the Negev.

[8] M. Last, et.al, , "Using Data Mining For Automated Software Testing", International Journal of Software Engineering and Knowledge Engineering, Vol. 14, No. 4 (2004) 369-393, World Scientific Publishing Company.

[9] Dr A.Kanagaraj, et.al, "Pervasive Computing Based Intelligent Energy Conservation System", Int. J. Advanced Networking and Applications, Volume: 07 Issue: 03 Pages: 2736-2740 (2015) ISSN: 0975-0290, IJANA.

[10] Amit Kumar et.al,, "Software Fault Prediction with Data Mining Techniques by Using Feature Selection Based Models", International Journal on Electrical Engineering and Informatics - Volume 10, Number 3, September 2018.

[11] Tao Xie, et al., "Mining Software Engineering Data", 29th International Conference on Software Engineering (ICSE'07 Companion), IEEE, Print ISBN: 0-7695-2892-9, 20-26 May 2007, Minneapolis, MN, USA.

[12] Mark Last, et al., "The data mining approach to automated software testing", SIGKDD '03, August 24-27, 2003, Washington, DC, USA, 2003 ACM 1-58113-737-0/03/0008.

[13] Mariam Bibi, et al., "Analytical Study of Data Mining Techniques for Software Quality Assurance", International Journal of Computer and Communication System Engineering (IJCCSE), Vol. 2 (3), 2015, 377-386, ISSN: 2312-7694.

[14] M. Halkidi, et al., "Data mining in software engineering", Intelligent Data Analysis 15 (2011) 413–441, DOI 10.3233/IDA20100475, IOS Press.

[15] Nidhin Thomas, et al., "Data Mining Techniques used in Software Engineering: A Survey", International Journal of Computer Sciences and Engineering, Volume-4, Issue-3, E-ISSN: 2347-2693, 2015, pp.28-34.

[16] Ali Ilkhani et.al,, "Extracting Test Cases by Using Data Mining; Reducing the Cost of Testing", International Journal of Computer Information Systems and Industrial Management Applications, ISSN 2150-7988 Volume 3 (2011) pp. 730-737 .

[17] Dr. R. Durga , Anulekshmi. s, Comprehensive Survey for Wireless Sensor Network and Internet of Things in Precision Agriculture", Jour of Adv Research in Dynamical & Control Systems, Vol. 12, No. 4, ISSN 1943-023X, Mar 2020, DOI: 10.5373/JARDCS/V12I4/20201427.

[18] Dr.R. Durga , P.Tamilselvi, "Detailed Review on Different Encryption Standards on Improved Cloud Data Security" , Jour of Adv Research in Dynamical & Control Systems, Vol. 12, No. 4, ISSN 1943-023X, Mar 2020, DOI: 10.5373/JARDCS/V12I4/20201428.

[19] Dr.Saradha , Kalaivani, Dr.R.Durga "Predicting an Interactions of Skin Lesion Using Neural Network", Journal of Information and Computational Science ,Volume 10 Issue 4 – 2020, ISSN: 1548-7741.

[20] R.Durga , G. Thailambal, P.Shanmugalakshmi A Novel Method of Rainfall Prediction using MLP-FFN and Hybrid Neural Network Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-10, August 2019.

[21] R.Durga, M.Prem Kumar , Analysis and Research on Integrated Multi Model Wireless Sensor AdHoc Network in Embedded Tracking Technology, JASC: Journal of Applied Science and Computations Volume V, Issue XII, December/2018, ISSN NO: 1076-5131.

[22] Dr.R. Durga, P.Manivannan, " Detection and Classification of Chronic Wounds Using Image Processing, Uploaded in Scopus, Feb 2020.

[23] Dr.R. Durga ,Zafer Ahmed N, "A survey on Intrusion Detection Schemes in Cloud Environment", Jour of Adv Research in Dynamical & Control Systems, Vol. 12, No. 4, ISSN 1943-023X, Mar 2020, DOI: 10.5373/JARDCS/V12I4/20201430.