Default Prediction in Software Module by using Support Vector Machine (SVM)

Meenakshi Sharma
Associate Professor(C.S.E)
S.S.C.E.T, Badhani
Pathankot, India .
Sharma.minaxi@gmail.com

Mona Pathania
Research Scholar(C.S.E)
S.S.C.E.T, Badhani
Pathankot, India
pathaniamona786@gmail.com

***Abstract*- Support Vector machine is the most popular and strong tool for machine learning. It has ability to evaluate maximum margin. As both faults and the failure are costly impact Therefore, in software development life cycle, it is necessary to predict defective modules in the early stage so as to improve and enhance software developer ability to identify the defect-prone modules and focus quality assurance activities.**

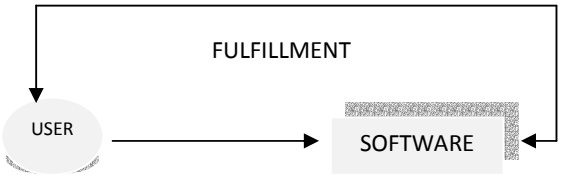
Keywords- Defect Prediction, Mining paradigms ,Support Vector machine.

I. INTRODUCTION

Faults and failures are costly circumstances. It is beneficial to predict the software that is a defect- prone. As quality and faults, handlings are the two factors that govern the cost of the software. Methods involve in maintaining the quality of software are significant, so the cost of this method can't be cut off. For cost optimization, the only way to reduce its cost is decreasing the cost for fault handling by adopting proper fault identification method. The attribute of the software can be measured with the different features such as complexity, pattern intricacy, endeavor, time mensuration, the dimension of the catalog, operands, operators, line count etc. There are many studies and learning passages that are used to calculate the performance metrics of the software. Scrutiny of all required features of defect prediction is used to resolve that what factors command predictive performance.

A. Software Engineering: Introduction

Software engineering is designated as the systematic and well-defined path to the development, operation, maintenance and evaluation of the software. This integrates the various preferred approaches to design the software which takes into consideration that what types of the machine the software will be used on, how the software will be work with the machine and what elements need to be put in place to ensure the reliability.



[12]

B. Attributes in Software Quality

In general, the quality attributes are the specification of non-functional requirement such as which comprises of following aspects. Traceability abstract deals with the accessibility to check the processing of software's internal functions according to the need so as to review its faults and problems meant as interoperability testing. The main motive of Efficiency in software quality is to check the structure and the composition of the system and also maintains so that all the resource can be utilized without any wastage. The flexibility of software is the adaptation of further changes occurs in an external environment. Next comes the availability which defines is the performance measurement of software quality so as the problems and faults do not affect its working. Disaster recovery leads the software tendency to keep on working and perform functioning. Security checks the protection of data in software which prevents it from any unauthorized access. Maintainability is an important section of quality attributes with the help of which systems improvements, up gradation, correction and enhancement take place so that

performance can be increased with minimum downtime.

QUALITY ATTRIBUTES
)81&7,21$/,7<#
5(/,$%,/,7<#
86$%,/,7<#
()),&,(1&<#
3257$%,/,7<#
0$,17$1(1&(#

[29]

C. Data Mining Paradigm

Data mining is the very persuasive approach for reducing information overwhelm and enhance decision making by squeezing and elevating useful knowledge from considerable dataset through a process of searching for relationships and patterns. Data mining is intertwined with the empiricism of neoteric and provocative patterns from the comprehensive. Data mining is described as the agreement of past and ongoing or recent evolution in stat artificial intelligence and machine learning. It is a crucial course in today's world because it divulges the obscure patterns for evaluation. These patterns can then be used for retailing scrutiny, making strategies, taking decisions; to increase wealth etc. Data mining gives a number of scientific tools for interpreting data. It provides various functionalities to data like multidimensional views of data, preprocessing of data, classifying data into classes according to their features, clustering the data etc. Scientifically, data mining is the process of finding interrelationships or arrangement among dozens of the area in huge databases.

II. RELATED WORK

There are many researchers who had worked on the default prediction of software module by using their technologies and learning approaches. In this paper, the major focus is to reduce the complexity of processing by feature extraction method and boundary condition problem which is resolved as Support Vector Machine (SVM) and component learning by using this method of adaptive boost with SVM and Radial Basis Function Kernel. Their work is reviewed and defined here.

David Gray et al. (2011) states that how we can make improvements in the efforts made in the project as well as to predict the fault of the software module. The method which is superior and useful in reducing the cost and defects is removed that is ignoring of the useless functionality by setting a line of code to a minimum.

Qinbao Song et al. (2011) provides up with the overall architecture and work that describes the evaluation schemes and components of fault prediction. For the given historical data sets the performance is measured and analyzed using scheme evaluation. Construction of models by fault predictors on the basis of evaluated learning scheme and faults of software are produced with new data so it concludes that for every data values there should be an use of different schemes.

Ming Li et al. (2012) showed that software fault prediction can control the quality of software. Nowadays maximum numbers of fault prediction methods are dependent on the huge amount of the historical data but if we talk about current and new projects then, in that case, historical data is unavailable for an organization. So to overcome this new sample-based method can be used by choosing and testing a less percentage of modules and then fault prediction models are used afterward to predict the various faults.

Martin Shepperd et al. (2014) focuses on the factors which have the highest effect on the performance of software by an analysis of the related studies on software module. Results described that the major problem is with the researcher group rather than choosing of classifier on the software performance

G Czibula et al. (2014) have discussed the importance of time software evolution and maintenance for the problem in fault prediction. By revealing the defective software module, the quality of the software is enhanced.

An Okutan et al. (2014) describes how different software metrics are being used for fault prediction and states the sets of metrics which is the major part for the prediction of efficiency in software module

Punika Mah. (2015) used the concept of WEKA and Mat lab for the collection of all the faults during the software development.

Prediction of the software faults is depicted which are based on the data mining.

III. SUPPORT VECTOR MACHINES

Support vector machines (SVM) are kernel based training theorem introduced by Vapnik in year of 1995 using concept of Principle Structural Risk Minimization (SRM) which minimizes the generalization error, SVM provides the extensibility and applicability that are needed in a production quality data mining system. It is a kernel-based algorithm. A kernel is a function that converts the input data to a high-dimensional slot where the problem is solved. Kernel functions can be linear or nonlinear. Machine learning is a body of knowledge that delve into the framework and study of algorithms that can learn from the data. Machine training mechanism is the commixture of demography and artificial intelligence and is meticulously related to computational statistics. Machine learning helps in taking decisions depending on the abstract of the studied data using statistics and computing more progressive expert systems interrogative and algorithms to gain results. SVM model plays a vital role in machine learning approach with standardized learning models which are interrelated with the algorithms on which the data can be analyzed as well as recognitions of patterns takes place. On another point, these are used for the classification and regression analysis. We can take an example of points in space, which is represented as SVM model.

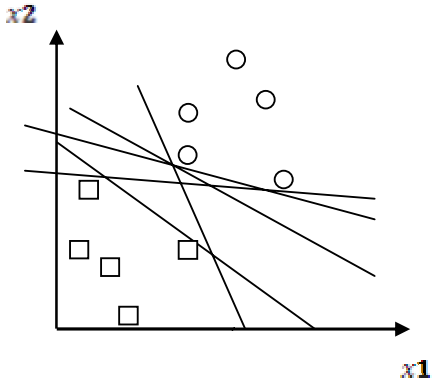


Fig 3.1 Hyper-plane separating two classes [12]

These points are mapped into various categories are the linear kernel, quadratic, polynomial, radial basis function, multi-perception.

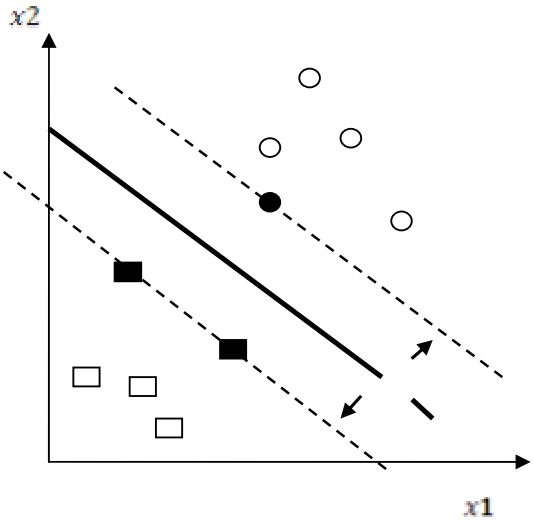


Fig 3.2 Optimal Hyperplane [12]

The given notation is used to describe a hyper plane:

$f(x) = \beta_0 + \beta^T x,$

Where β = *Weight vector* & β_0 = B*ias*

(i)

The numerator is equal to one and the distance to the support vectors, for the canonical hyperplane, is

$$distance_{support\ vector} = \frac{|\beta_0 + \beta^T x|}{||\beta||} = \frac{1}{||\beta||}$$

(ii)

Remember that the boundary values introduced in the above figure is represented as M here, and is twice the distance to the nearby examples:

$$M = \frac{2}{||\beta||}$$

Formally,

$$distance_{support\ vector} = \frac{|\beta_0 + \beta^T x|}{||\beta||} = \frac{1}{||\beta||}$$

(iii)

IV. EMPIRICAL EVALUATION

In this paper, the methods involve the analysis of classification with ranking features using machine learning approach that is support vector machine is discussed. To get a précised accurate observations we use training sets with SVM radial basis function, further all the steps

include in the classification of reduced sets of features, kernel approach is defined for linear and non-linear data description. There are many kernels such as RBF, polynomial etc for supreme results so well-defined sets of features are used through which information is extracted from parameters in the data mining.
Stage1: Description of suitable data values over the features like design complexity, effort, time mensuration, line count as complexity, pattern intricacy, endeavor, the dimension of the catalog, operands, operators, line count etc.
Stage2: Execution of features extraction method over the actual data values by the use of principal extraction method which is used to combine the data values.
Stage3: All the different features like P1, P2, P3…Pn are taken into an account and then it gives the status of the values, whether are they default not default that is {+1,-1}. Suppose if any one of the value comes +1 then it will count under default else, not default.
Stage4: Now adaptive boost with SVM-RBF kernel is implemented so as to split off the compaction and boundary error condition in the feature extraction method.
Stage5: Last step is to find out the accuracy, recall and the precision of the software module By applying classifier model of SVM machine.

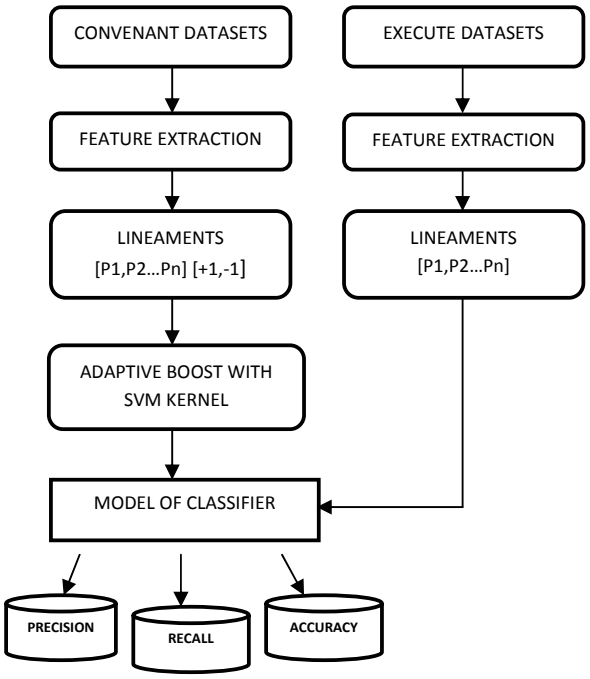


Fig 4.1Methodology of Proposed Work [31]

- **Adaptive boost learning approach**

This machine learning methods can be used as a Meta implementation with all other. For the improvement in the performance, these can be used parallel with many other learning programs. Their results are summed up which are derived from those learning algorithms, At last, it gives a boosted classifier considered as a final output.

V. DISCUSSION OF RESULTS

Statistics are defined for the data values including many classifiers which are trained and executed using the label Train and Test. Below given Table presents the SVM parameters using classifiers which provide the actual percentage value for accuracy, recall and precision sets. This whole examining of parameters is done with support vector machine model using RBF kernel.

Classifier	*Precision*	*Recall*	*Accuracy*
Linear	71.31	72.89	77.78
Quadratic	65.41	68.94	72.88
Polynomial	65.96	75.44	72.31
RBF	55.64	64.35	66.20
Multilayer perception	54.96	55.33	66.67

Table 5.1 Analysis Parameters of Precision, Recall and Accuracy on different SVM with 12 features.

Here is the depiction of confusion matrix used with the help of machine learning approach. Every row of confusion matrix shows instances in certain class while other column shows instances in presumed class.

Sensitivity	*Specificity*	
255	124	***Sensitivity***
347	876	***Specificity***

Table 5.2 Confusion Matrix

VI. OVERALL OBSERVATIONS

- Precision provides an overall fraction of the instances retrieved which are consistent.

Precision=Total consistent data/Total Data extracted

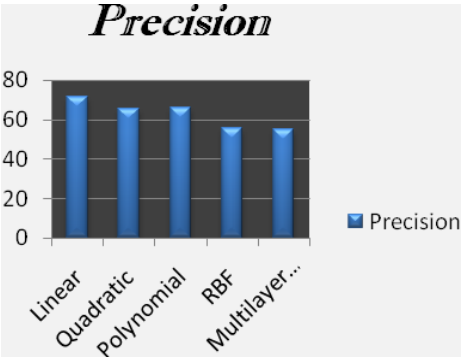


Fig 6.1- Comparison of Precision of software model with different SVM's

- Recall gives the overall fraction values of retrieved exponents which have similar behavior.

Recall=Total Data Extracted/Total Related Data.

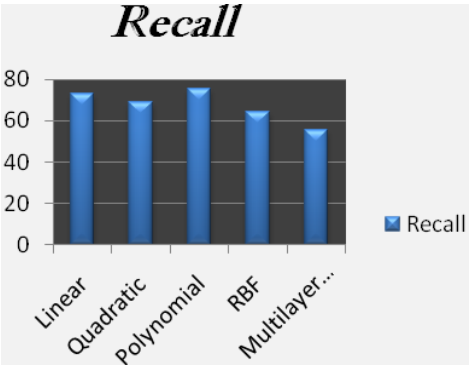


Fig 6.2- Comparison of Recall of Software Model with different SVM's

- Accuracy is one of the visceral performance mensuration which is the ratio of the precisely gathered perceptions.

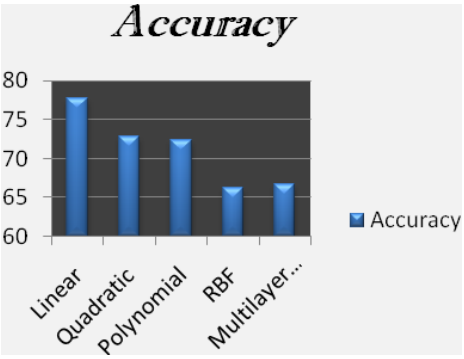


Fig 6.3-Comparison of Accuracy of Software Model with different SVM's

- Comparison of SVM machine over class 0 class 1 class 2 class 3 are described below where 0 indicates to the Linear SVM machine, 1 indicates Quadratic set, 2 indicates as a polynomial, 3 indicates RBF kernel and last classifier 4 represents that class is of Multilayer perception. All the results of SVM are compared in the graph which provides with improvement. This is based on RBF kernel giving optimal instances for parameters.

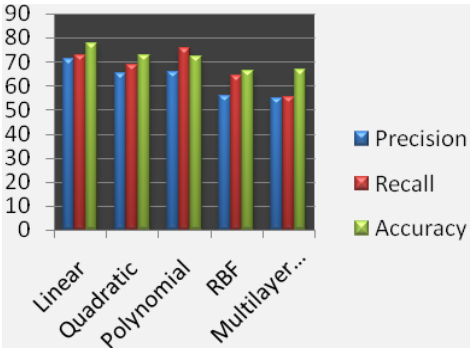


Fig 6.4-Complete comparison graph of Parameters with different SVM's

By correlating the outcomes from the table and graphs we achieved a maximum prediction in good form, so this feature of SVM machine is helping in attaining a better selection and understanding of the information.

VII. CONCLUSION AND FUTURE

SCOPE

Above tables and graphs clearly depicts that Adaptive Boost with SVM learning approach is the best approach for the prediction of default and not default features of software model as better precision, recall and accuracy is much better than all other learning approaches**.** Even by reducing the features from high to low quantity for analyzing the various parameters of the software model, SVM learning approach gives better precision, recall, and accuracy than the other approaches.

Future Scope

The selected are can be chosen and applied on different modules of the machine for the enhancement in improvement, In future new hybrid techniques can be used with more parameters in machine learning.

REFERENCES

[1] N. Fenton and M. Neil, "A Critique of Software Defect Prediction Models," *IEEE Trans. Softw. Eng.*, vol. 25, no. 5, pp. 675–689, 2000.

[2] T. Zhang, "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods," *AI Mag.*, vol. 22, no. 2, p. 103, 2001.

[3] N. Cristianini and B. Schölkopf, "Vector Machines and Kernel Methods Learning Machines," *AI Mag.*, vol. 23, no. 3, pp. 31–42, 2002.

[4] T. Joachims,eatl, "Introduction to Support Vector Machines," *Distribution*, pp. 1–15, 2002.

[5] R. Berwick, "An Idiot's guide to Support vector machines (SVMs)," *Retrieved Oct.*, pp. 1–25, 2003.

[6] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel.," *Neural Comput.*, vol. 15, no. 7, pp. 1667–89, 2003.

[7] N. Cristianini and J. Shawe-Taylor, "Support vector and kernel methods," *Intell. data Anal.*, 2003

[8] H. Zhang, M. Genton, and P. Liu, "Compactly supported radial basis function kernels," *Available www4. stat. ncsu. edu/hzhang/ …*, pp. 1–21, 2004.

[9] F. Markowetz, "Classification by Support Vector Machines," 2004.

[10] M. Bhasin and G. P. S. Raghava, "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST," *Nucleic Acids Res.*, vol. 32, no. WEB SERVER ISS., pp. 414–419, 2004.

[11] G. Casella, S. Fienberg, and I. Olkin, *Springer Texts in Statistics*, vol. 102. 2006.

[12] K. O. Elish and M. O. Elish, "Predicting defect-prone software modules using support vector machines," *J. Syst. Softw.*, vol. 81, no. 5, pp. 649–660, 2008.

[13] H. Cao, T. Naito, and Y. Ninomiya, "Approximate RBF Kernel SVM and Its Applications in Pedestrian Classification," *Mach. Learn.*, pp. 1–9, 2008.

[14] X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[15] A. Smola and S. V. N. Vishwanathan, "Introduction to Machine Learning," 2008.

[16] R. A. Calix and R. Sankaran, "Feature Ranking and Support Vector Machines Classification Analysis of the NSL-KDD Intrusion Detection Corpus," no. 2006, pp. 292–295, 2009.

[17] L. Hamel, *Knowledge Discovery With Support Vector Machines*. 2009.

[18] S. Beecham, T. Hall, D. Bowes, D. Gray, S. Counsell, and S. Black, "A Systematic Review of Fault Prediction approaches used in Software

Engineering," *Engineering*, no. 03, 2010.

[19] Y. Singh, A. Kaur, and R. Malhotra, "Prediction of Fault-Prone Software Modules using Statistical and Machine Learning Methods," *Int. J. Comput. Appl.*, vol. 1, no. 22, pp. 8–15, 2010.

[20] X. Zhang, "Support vector machines," *Encycl. Mach. Learn.*, vol. 1, pp. 941–946, 2010.

[21] A. Ben-Hur and J. Weston, "A user's guide to support vector machines.," *Methods Mol. Biol.*, vol. 609, pp. 223–239, 2010.

[22] S. R. Das, P. K. Panigrahi, K. Das, and D. Mishra, "Improving RBF Kernel Function of Support Vector Machine using Particle Swarm Optimization," no. 4, pp. 130–135, 2012.

[23] A. Khajeh-Hosseini, D. Greenwood, J. Smith, and I. Sommerville, "The Cloud Adoption Toolkit: supporting cloud adoption decisions in the enterprise," *Softw. - Pract. Exp.*, vol. 43, no. 4, pp. 447–465, 2012.

[24] P. Banerjee, P. Banerjee, and S. S. Dhal, "International Journal of Advanced Research in Computer Science and Software Engineering," *Int. J.*, vol. 2, no. 9, pp. 62 – 70, 2012.

[25] M. Parviainen, "Radial Basis Function (RBF) and Support Vector Machines (SVM) networks Radial Basis Function (RBF) networks," 2012.

[26] T. Nadu, "Bagged SVM Classifier for Software Fault Prediction," vol. 62, no. 15, pp. 21–24, 2013.

[27] R. Huerta, F. Corbacho, and C. Elkan, "Nonlinear support vector machines can systematically identify stocks with high and low future returns," *Algorithmic Financ.*, vol. 2, pp. 45–58, 2013.

[28] P. Thangaraj and N. Renukadevi, "Performance Evaluation of SVM–RBF Kernel for Medical Image Classification," *Glob. J. Comput. …*, vol. 13, no. 4, 2013.

[29] A. Adline and M. Ramachandran, "Predicting the Software Fault Using the Method of," pp. 390–398, 2014.

[30] M. G. Feshki, "Managing Intrusion Detection Alerts Using Support Vector Machines," no. 9, pp. 266–273, 2015.

[31] P. Mahajan, "International Journal of Advanced Research in Computer Science and Software Engineering Improved Default Prediction in Software Module by using Feature Extraction and Adaptive-Boost Learning Approach," vol. 5, no. 5, pp. 524–528, 2015.