Analysis Receiver Operating Characteristics of Software Quality Requirement by Classification Algorithms

Dhyan Chandra Yadav
Research Scholar,
Shri Venkateshwara University,
Gajraula, Amroha (U.P.)

Saurabh Pal
Head, Dept. of MCA,
VBS Purvanchal University.
Jaunpur (U.P.)

ABSTRACT
Requirement engineering has an important role in software project development. Quality maintenance is the major factor in software industry. Requirement continues increases in the software market at different economic status with high class quality. The quality of software project development depend on technical performance but generally a technical problem run in project development known as duplicity. Duplicity is software bug which create problem in development. Data mining generate technical help in analysis of problematic area. In this paper we proposed the analysis of receiver operating characteristics of software defect related attribute data object and also analysis cost/benefit population, target confusion matrix and classification accuracy by zeroR, oneR and Prism algorithms of data mining.

General Terms
Data Mining, Classification algorithms, Software Engineering, Weka Tool.

Keywords
Data Mining; Classification: zeroR ,oneR and Prism; ROC; Weka.

1. INTRODUCTION
Williams [1] discussed that duplicity is a technical problem in software project development life cycle. Duplicity is closed as more spaces between codes, change code in another location and error down line. All the technical problems mentation in a report known as problems report .If any bug reported in a problem report but it is already covered by another problem report this happening is known as duplicity of bug. Duplicate bug is created in any phase testing and it is some time automatically created in the coding implementation or phase testing by the help of data mining. It is easily classified and analyzed in the software engineering domain.

Tiwari and Chaudhary **[2]** introduced about Data mining. Data mining provide facility in analysing data from different useful information. For example: Data Mining analysed bug or no bug in software project from related information. Data mining classifying the problems by zeroR, oneR and Prism algorithms and get accuracy of related attributes.

Holte [3] discussed ZERO-R is an in consequential classifier, but it provides a minor certain presentation of a software defect database which should be suggestively better quality of software. It provides a sensible test on how glowing bug/no bug class can be expected without bearing in mind the other attributes. It is work as a lower bound for software defect data set and checks the instances values with class prediction. zeroR categorized numeric prediction problem in training data.

In Weka [4] oneR is the second simplest classification model. It is planned for insignificant facts. It products modest instructions grounded on a particular quality. It generates one level decision tree for software defect and also count how each software defect class appears. OneR makes a rule for each Software defect attribute in training data and also calculates the error rate by this rule.

Hong and Tseng [5] apply prism algorithm which has the impression of evidence improvement in its place of entropy as ID3. Quality respected couples in relationships of information theory, can be supposed of separate communications.

The quantity of evidence improvement about a happening in a communication is defined as:

Evidence improvement is selected for recounting a class with a larger importance.

The mission of the prism algorithm is:

a) Calculate the probability of incidence of the arrangement for each chooser.

b) Choose the chooser for which probability incidence is a maximum then create a subset of the preparation set

c) Repeat step 1 and 2 for this subset until it contains only instances of class classification.

d) The multifaceted law is combination of all the choosers used in creating the comparable sub selection.

e) At implementation set, remove all instance enclosed by multifaceted law.

f) Reprat steps 1-5 until all occurrences of class classification have been detached.

Swets and John [6] discussed that in figure a receiver operating characteristic (ROC) is a graphical plot that demonstrates the presentation of a binary classifier scheme as its judgment threshold is diverse. The arc is created by trickery the true positive rate beside the false positive rate at many threshold settings. The true positive rate is also known as sensitivity or recall in machine learning. The false positive rate is known as the fall out and can be calculated as specificity. The ROC curve is thus the sensitive as a function of fall out. In general, if the probability circulations for both discovery and false fear are known the ROC curve can be produced by trickery the cumulative distribution function of the detection probability in the y-axis versus the growing circulation function of the false arm possibility in x-axis.

2. RELATED WORK

Shepperd, Schofield and Kitchenham [7] discussed that need of cost estimation for management and software development organizations and give the idea of prediction also give the methods for estimation.

Alsmadi and Magel [8] discussed that how data mining provide facility in new software project its quality, cost and complexity also build a channel between data mining and software engineering.

Boehm, Clark, Horowitz, Madachy, Shelby and Westland [9] discussed that some software companies suffer from some accuracy problems depend on his data set after prediction software company provide new idea to specify project cost schedule and determine staff time table.

Chaurasia and Pal [10, 11] conducted study on the prediction of heart attack risk levels from the heart disease database with data mining technique like Naïve Bayes, J48 decision tree and Bagging approaches and CART, ID3 and Decision Table. The outcome shows that bagging techniques performance is more accurate than Bayesian classification and J48.

Ribu [12] discussed that the need of open source code projects analyzed by prediction and get estimating object oriented software project by case model.

Nagwani and Verma [13] discussed that the prediction of software defect(bug) and duration similar bug and bug average in all software summery, by data mining also discuss about software bug.

Hassan [14] discussed that the complex data source(audio, video, text etc.) need more of buffer for processing it does not support general size and length of buffer.

Li and Reformate [15] discussed that .the software configuration management a system includes documents, software code, status accounting, design model defect tracking and also include revision data.

Pal and Pal [16] conducted study on the student performance based by selecting 200 students from BCA course. By means of ID3, c4.5 and Bagging they find that SSG, HSG, Focc, Fqual and FAIn were highly correlated with the student academic performance.

Elcan [17] discussed that COCOMO model pruned accurate cost estimation and there are many thing about cost estimation because in project development involve more variable so COCOMO measure in term effort and metrics.

Chang and Chu [18] discussed that for discovering pattern of large database and its variables also relation between them by association rule of data mining.

Kotsiantis and Kanellopoulos [19] discussed that high severity defect in software project development and also discussed the pattern provide facility in prediction and associative rule reducing number of pass in database.

Pal [20] conducted study on the student dropout rate by selecting 1650 students from different branches of engineering college. In their study, it was found that student's dropout rate in engineering exam, high school grade; senior secondary exam grade, family annual income and mother's occupation were highly correlated with the student academic performance.

Pannurat, Kerdprasop and Kerdprasop [21] discussed that association rule provide facility the relationship among large dataset as like software project term hug amount , cost record and helpful in process of project development.

Fayyad, Shapiro, Smuth and Uthurusamy [22] discussed that classification creates a relationship or map between data item and predefined classes.

Shtern and Vassillios [23] discussed that in clustering analysis the similar object placed in the same cluster also sorting attribute into group so that the variation between clusters is maximized relative to variation within clusters.

Runeson and Nyholm [24] discussed that code duplication is a problem which is language independent. It is appear again and again another problem report in software development and duplication arises using neural language with data mining.

Vishal and Gurpreet [25] discussed that data mining analyzing information and research of hidden information from the text in software project development.

Yadav and Pal [26] conducted a study using classification tree to predict student academic performance using students' gender, admission type, previous schools marks, medium of teaching, location of living, accommodation type, father's qualification, mother's qualification, father's occupation, mother's occupation, family annual income and so on. In their study, they achieved around 62.22%, 62.22% and 67.77% overall prediction accuracy using ID3, CART and C4.5 decision tree algorithms respectively.

3. METHODOLOGY

3.1 Data Preparation

A software error arises in problem report and all problem reports grouped in two categories: recoverable and unrecoverable. In recoverable group an error easily recovered automatically by software bug tracking system GANTS. It is set up on MASC intranet to collect and maintain all problem reports from every department of MASC. If the bug is already covered by another problem report is known as duplicate-bug. Duplicate-bug arises in code implementation. Now performing for classification of duplicate-bug using several standard data mining tasks, data preprocessing, clustering, classification, association and tasks are needed to be done. The database is designed in MS-Excel, MS Access 2010 database. The data is formed according to the required format and structures and data is converted to ARFF (attribute relation file format) format to process in Weka. An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes.

3.2 Data Selection and Transformation

In this study, we set up a series of classification experiments focusing three algorithms in Weka, data mining tool which are zeroR, oneR and prism. The task is to predict and measure if SRS has been reused information from Software Company. For the given data set the collection of attribute value from duplicate bug that belong to the bug group forms the distribution which presents the number of bugs. The experiments were carried out in two stages. First stage is to measure the reuse performance of zeroR, oneR and prism. The roc areas were observed and recorded. Next in the second stage measure cost/benefit added to the experiments.

Table 1.The variables used in the computational technique

Property	Description	
Source	Name of a project or department that rises the PR.	
Measurement Type	(Duplicate –BUG) SRS with metrics count	
Sample Size	61 Total: 19 duplicate-BUG and 42 Non duplicate BUG	
Dependable Variables		
Property	**Description**	**Possible Values**
Unambiguous	To obtain the percentage of requirements that has been interpreted in a unique manner by all its reviewers.	0-Ambigugous requirements 1-Unambiguous
Correct	To measure percentage of requirements in the SRS that has been validates.	0-incorrect 1-correct
Complete	To measure total number of functions currently specified.	Grows closer to 1.
Verifiable	To measure the verifiability of a requirement.	0- very Poor 1- Very Good
Traceable	To measure which requirements are being supported by the components or verified by testing? If a single requirement is not traceable then the whole SRS is untraceable.	1-traced attribute 0-otherwise
Not Redundant	To measure the percentage of unique functions that is not repeated.	1-No redundancy 0-complete redundant
Design Independent	To meet the percentage of possible solution system that are eliminated by adding the overly constraining requirements.	1-design Independent 0-highly design independent

3.3 Data Mining Implementation

In this step, three classification algorithms are chosen for the purpose of accuracy in dataset, which are the zeroR, oneR and prism. To investigate further the classifier performance in accuracy. The ROC graph organizes classifiers and helps visualize their performance. ROC graph are commonly used in software project development decision making. ROC is basically two dimensional graphs in which true positive or legal is plotted on to y-axis and false positive or bug is plotted on the x-axis. The classifier that is nearest to the perfect or the top left corner in the graph shows the best accuracy.

4. RESULT AND DISCUSSION

4.1 Experimemt-1

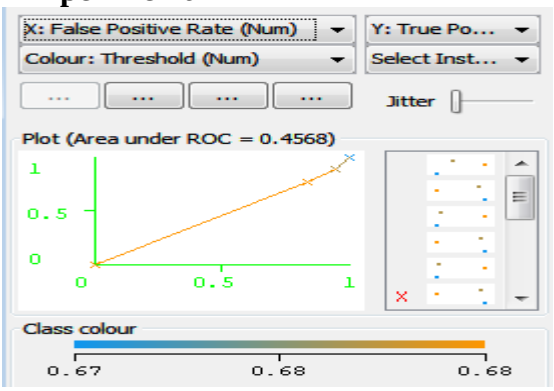

Fig 1: Representation of zeroR in Weka

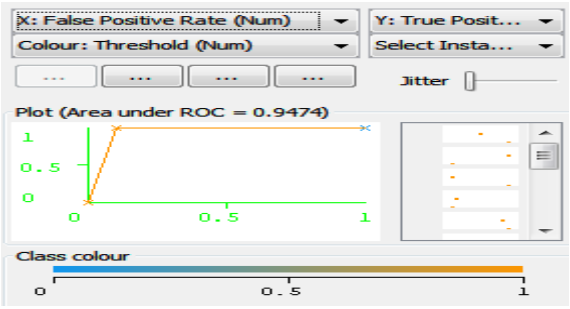

Fig 2: Representation of prism in Weka.

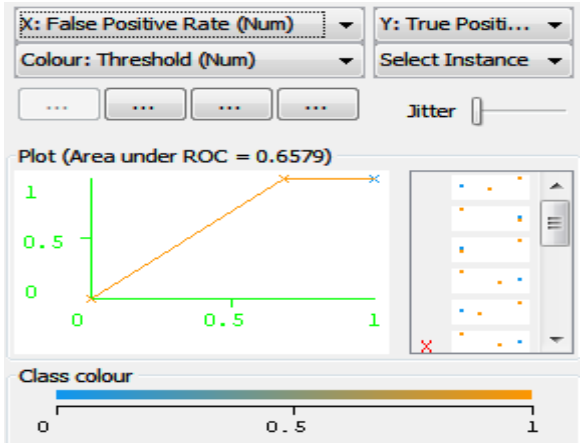

Fig 3: Representation of oneR in Weka.

Table 2. ROC Representatio

ALGORITHMS	ROC
zeroR	0.4568
Prism	0.9474
oneR	0.6579

The experiment results are reported in three parts algorithmic (zeroR, oneR, Prism) selection is applied. Results from observation, of reusable ROC area using zeroR 0.4568 whereas 0.9474 and 0.6579 for prism and oneR respectively.

Basically, ROC is a two dimensional graph in which true positive is plotted on the y-axis and false positive is plotted on the x-axis.

The classifier that is nearest to the perfect point (0, 1) or top left corner in the graph shows the best accuracy. The ROC areas serve to present the comparative performance across three proposed classifier the performance of the binary classifier.

In this study we set up a series of classification experiments that the prism shows the best result in terms of accuracy in our experiment.

4.2 Experiment-2

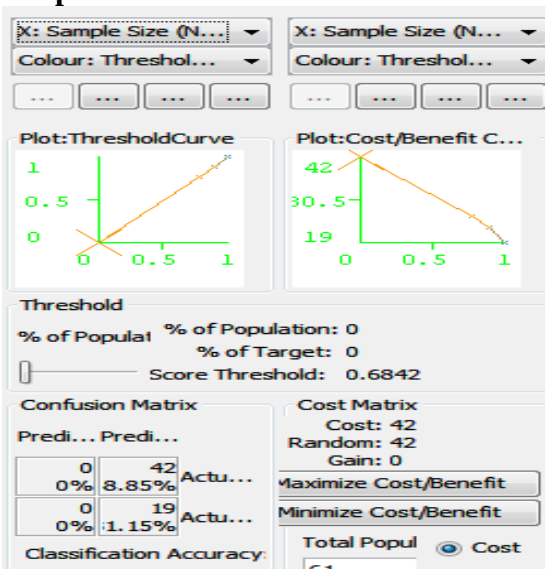

Fig 4: Representation of cost/ benefit curve of zeroR

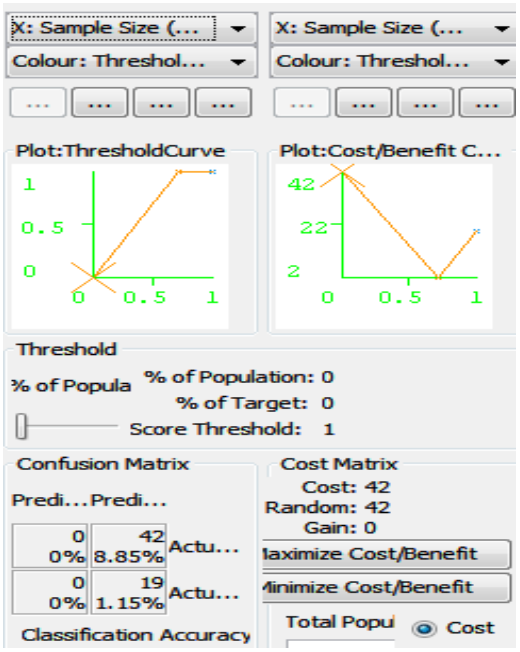

Fig 5: Representation of cost/ benefit curve of prism

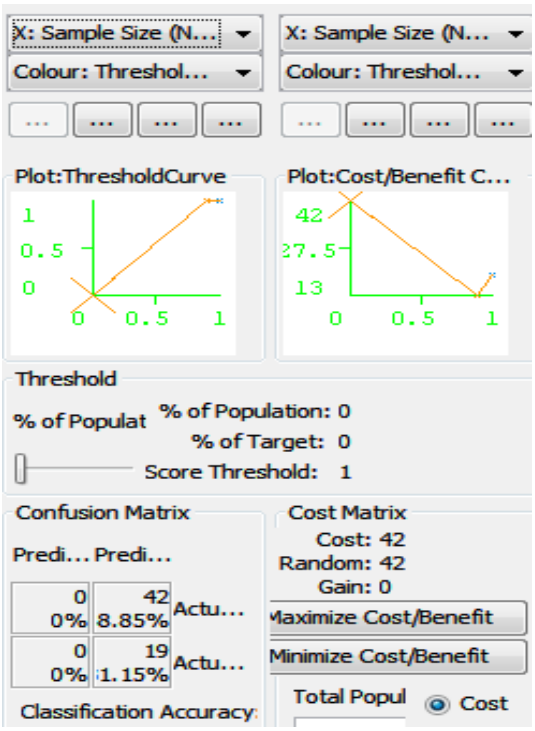

Fig 6: Representation of cost/ benefit curve of oneR

Table 3. Representation of Cost/Benefit Details

Detailed	Iteration	zeroR	Prism	oneR
Correctly Classified	1	31.15	31.15	31.15
	2	57.38	96.72	78.69
	3	62.30	68.82	68.85
	4	68.85		
Score Threshold	1	0.68	1	1
	2	0.68	1	1
	3	0.68	0	0
	4	0.67		
Cost/benefit	1	42	42	42
	2	26	2	13
	3	23	19	19
	4	19		
Random	1	42	42	42
	2	23.9	25.41	21.26
	3	21.26	19	19

	4	19		
Recall(Target)	1	0	0	0
	2	76.19	100	100
	3	88.10	100	100
	4	100		
Gain	1	0	0	0
	2	-2.1	23.41	8.26
	3	-1.74	0	0
	4	0		
Population	1	0	0	0
	2	78.69	72.13	90.16
	3	90.16	100	100
	4	100		

ROC analysis is related in direct and natural way to cost/benefit analysis of bugs by tool which represents chart on the left gain and a right gain is a commutative gains. A slider is provided to allow the user to explore the cost/benefit associated with various subsets of the population, various level of recall (percentage of target) or various thresholds on the probability of predicting the positive class. From the experiment we find average correctly classified of different iterations it is clear that prism give the correctly classified compare to zeroR and oneR algorithms.

It has been observed that the prism shows the best result in terms of accuracy in our experiment.

5. CONCLUSION

In present study we have used three different classification algorithms in a data mining tool, Weka using standard quality of software data sets and compared the accuracy level of each method. It has been observed that the prism shows the best result in term of accuracy in our experiment. However, we find prism also give more accurate result compares to zeroR and oneR algorithms. In the future work, we hope to investigate further on attributes from other software dataset.

6. REFERENCES

[1] Williams A., "Database Tip: Eliminate Duplicate Data" Friday 25 January 2008.

[2] Tiwari S. and Chaudhary N., "Data mining And Warehousing" Dhanpati Rai and Co.(P) Ltd. First edition: 2010.

[3] Holte, R.C., 1993 Very simple classification rules Perform well on most commonly used datasets. Machine Learning Vol 11, pp 63-91.

[4] OneR:http://en.wikipedia.org/wiki/One-attribute-rule 16 April 2007.

[5] Tzung-Pei Hong and Shian Shyong, "Tseng; Two-phase PRISM Learning Algorithms", Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, IEEE International Conference, Vol. 4, pp 3895 – 3899,1997

[6] Swets, John A., "Signal detection theory and ROC analysis in psychology and diagnostics", collected papers, Lawrence Erlbaum Associates, Mahwah, NJ, 1996.

[7] Shepperd M., Schofield C. and Kitchenham B., "Effort estimation using analogy," in of the 18th International Conference On Software Engineering, pp.170- 178. Berlin, Germany, 1996.

[8] Alsmadi and Magel, "Open source evolution Analysis," in proceeding of the 22nd IEEE International Conference on Software Maintenance (ICMS'06), phladelphia, pa.USA, 2006.

[9] Boehm, Clark, Horowitz, Madachy, Shelbyand Westland, "Cost models for future software life cycle Process COCOMO2.0.", in Annals of software Engineering special volume on software process and product measurement, J.D.Arther and S.M.Henry, Eds, vol.1, pp.45-60, j.c. Baltzer AG,science publishers, Amsterdam,The Netherlnds, 1995.

[10] Chauraisa V. and Pal S., "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.

[11] Chauraisa V. and Pal S., "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j.SciTech,,Vol.1, pp. 208-217, 2013.

[12] Ribu,Estimating "Object oriented software projects With use cases", M.S. thesis, University of Oslo Department of informatics, 2001.

[13] Nagwani N. and Verma S., "Prediction data mining Model for software bug estimation using average Weighted similiarity," In proceeding of advance Computing conference (IACC), 2010.

[14] Hassan A. E., "The road ahead for mining software Repositories", in processing of the future of software Maintenance at the 24th IEEE international Conference on software maintenance, 2008.

[15] Li Z. and Reformat, "A practical method for the Software fault prediction", in proceeding of IEEE Nation conference information reuse and Integration (IRI), 2007.

[16] Pal A. K., and Pal S., "Analysis and Mining of Educational Data for Predicting the Performance of Students", (IJECCE) International Journal of Electronics Communication and Computer Engineering, Vol. 4, Issue 5, pp. 1560-1565, ISSN: 2278-4209, 2013.

[17] Elcan C., "The foundations of cost sensitive learning", In proceeding of the 17 International conference on Machine learning, 2001.

[18] Chang C. and Chu C., "software defect prediction Using international association rule mining", 2009.

[19] Kotsiantis and Kanellopoulos, "Associan rule mining: A recent overview", GESTS international transaction on computer science and Engineering, 2006.

[20] Pal S., "Mining Educational Data to Reduce Dropout Rates of Engineering Students", I.J. Information Engineering and Electronic Business (IJIEEB), Vol. 4, No. 2, 2012, pp. 1-7.

[21] Pannurat, Kerdprasop and Kerdprasop, "Database reverses engineering based On Association rule mining", IJCSI, 2010.

[22] Fayyad, Piatesky Shapiro, Smuth and Uthurusamy, "Advances in knowledge discovery and Data Mining", AAAI Press, 1996.

[23] Shtern M. and Vassilios, "Review article advances in Software engineering clustering methodologies for Software engineering", Tzerpos volume, 2012.

[24] Runeson P. and Nyholm O., "Detection of duplicate Defect report using neural network processing", in Proceeding of the 29th international conference on Software engineering 2007.

[25] Vishal G. and Gurpreet S. L., "A survey of text mining Techniques and applications", journal of engineering Technologies in web intelligence, 2009.

[26] Yadav S. K. and Pal S., "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.