Vol 8. No. 2 June, 2015

African Journal of Computing & ICT

www.ajocict.net

An Integration of Clustering and Classification Technique in Software Error Detection

D.C. Yadav
Research Scholar,
Shri Venkatwhwara University, Amroha
E-mail: dc9532105114@gmail.com

Pal, S.
Department of MCA,
VBS Purvanchal University Jaunpur (U.P.)
E-mail: drsaurabhpal@yahoo.co.in

ABSTRACT

The software project development plays important role in software quality. Measuring software quality in a specific manner such as error estimation and check severity of error which is related to the document bug .Data mining create a supportable platform for software project development by which software engineers easily achieved goal of project quality in given time duration and budget .In this paper we integrate both technique classification and clustering for software error detection. Classification technique analyzes the severity of software defect by J48GRAFT, LAD Tree, and BAYESNET also by Clustering technique measure maximum similar data object in data set within same cluster by K-Means.

Keywords - Data mining; Classification: J48graft, Lad Tree and BayesNet; Clustering: K-Means; Weka.

African Journal of Computing & ICT Reference Format:
D.C. Yadav & S. Pal (2015). An Integration of Clustering and Classification Technique In Software Error Detection.
Afr J. of Comp & ICTs. Vol 8, No. 2. Pp9-16.

I. INTRODUCTION

Software testing check the performance of software project in which, software tracker plays a technical role to detect the software defect. The software tracker provides feedback information about bug. And give the technical role in software quality improvement. Software tracker provides the severity of bug fix or not fix in the software. Tracker measuring software quality in a specific manner such as error estimation and check the severity of error which is related to the document is known as document bug also decide the severity of bug. Shi and Harjan [1] in 2007 presented data classification of software defect is very commanding mission in data mining. The categories of software defects in data mining used comparable tree function, rule etc. The purpose of arrangement is to analyze the ideas of each adjustable and allocate those variables to corresponding predefined classes.

Jachyra, Pancerz and Gomula [2] introduced about J48graft. J48graft generates a technical graphical way to describe unrecovered problems of document bug in training set. The J48graft generates decision tree and give the training set but it have some drawback which is recover by J48graft. J48graft classify the environment into multi-dimensional space which is not possible by training set it is also reduce the prediction error. In the Fig. 1 J48graft classify the bug of software defect and easily we analyzed bug by J48graft tree.

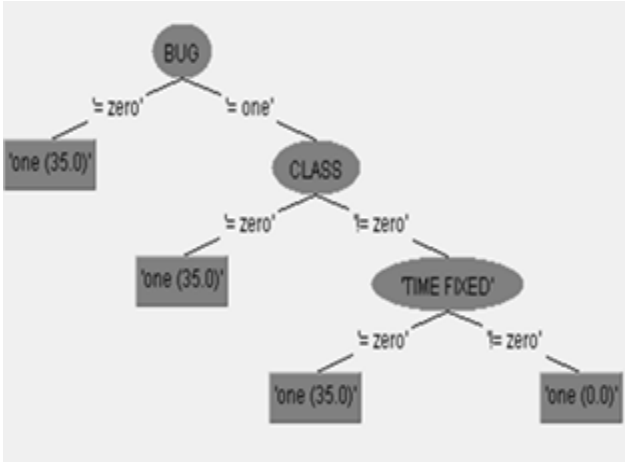

Figure 1. J48graft tree represents attribute classification of doc-bug using Weka.

Buhmann [3] introduced about Lad tree. Lad tree support in the logical analysis of data. It builds a classifier for binary target variable based on learning a logical expression that can distinguish between positive and negative sample in a data set.

LAD tree analyzed software defect in term of binary but not by any negative pattern is positive and similar; binary points covered by some negative pattern but not covered by positive pattern is negative.

Figure 2. Lad tree represents attribute classification of doc-bug using Weka.

The selection of software defect data set, sub set and construct LAD tree model which satisfied the above assumption such each pattern in the model. Pai and Dugan [4] introduced about Bayes Net and calculate the condition probabilities between variables.

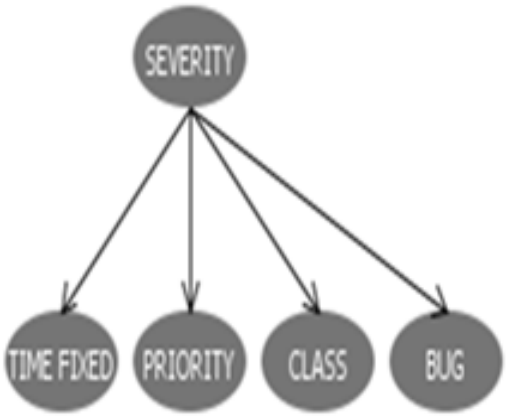

Figure 3. Bayes Net represents attribute classification of doc-bug using Weka.

Bayes Net give join probability distribution as a directed a cycle graph and local probability distribution. SunitaTiwari and NehaChaudhary [5] introduced about clustering. Clustering is a composition of data in classes but does not known the class level of all classes. Clustering has a minor different compare with classification and classification have surely class level conform to training set so clustering is unsupervised classification.

Clustering is depending on maximum the similarity between object in same class and minimizing the similarity objects of different classes.Tiwary and Chaudhary introduced about K-means algorithm. It is a centroid based partitions or cluster. K-means algorithm provide k- cluster on the data set.

Let d_1, d_2…….d_n. are data points of software defect data set. Let C_i denotes cluster number for the ith data point. K-means minimize scatter by

$$M(c)=1/2 \sum_{k=1}^{k} \sum_{c(i)=k} \sum_{c(j)=k} \| di-dj \|^2$$

$$= \sum_{K=1}^{K} N(k) \sum_{c(i)} \| d(i)- c(k) \|^2 \qquad \text{……...(1)}$$

When C_k is the mean vector of the Kth cluster and N_k is the number of data point in Kth cluster.

2. RELATED WORKS

Shepperd, Schofield and Kitchenham [6] discussed that need of cost estimation for management and software development organizations and give the idea of prediction also give the methods for estimation. Yadav and Pal [7] use the ID3 decision tree to generate the important rules that can help to predict student enrollment into an academic programme called the Master of Computer Application. The generated tree yields that Bachelor of Science students in mathematics and computer applications will enroll and will likely to perform better as compared to Bachelor of Science students without any background in mathematics.

Pal and Pal [8] 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. Alsmadi and Magel [9] 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 et al. [10] 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. Ribu [11] 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 [12] 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 [13] 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.Yadav and Pal [14, 15] discussed the use of different classification algorithms using standard quality of software data sets and compared the accuracy level of each method. Li and Reformate [16] discussed that .the software configuration management a system includes documents, software code, status accounting, design model defect tracking and also include revision data.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. Pannurat, Kerdprasop and Kerdprasop [20] 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, Piatesky Shapiro, Smuth and Uthurusamy [21] discussed that classification creates a relationship or map between data item and predefined classes. Shtern and Vassillios [22] 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 [23] 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 [24] discussed that data mining analyzing information and research of hidden information from the text in software project development.Lovedeep and Arti [25]data mining provide a specific platform for software engineering in which many task run easily with best quality and reduce the cost and high profile problems.Nayak and Qiu [26] discussed that generally time and cost, related problems arises in software project development these problems mentation in problem report ,data mining provide help in to reduce problems also classify and reduce another software related bugs .

Chaurasia and Pal [27, 28] 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.

1) The proposed system will analyze severity of software defects predicts. Predicts categorical class level classifiers based on training set and the values in the class level attribute use the model in classifying new data. We integrate both (classification and clustering) techniques. After combine application of most frequent used clustering (k-means) algorithm with classification (J48GRAFT, LAD TREE and BAYESNET) algorithms, the results were compare and the weka data mining tool was used.

3. METHODOLOGY

Our research approach is to use J48graft, Lad Tree, Bayes Net and K-Means; to model the relationships between themeasurable properties of a software product and its quality.The research methodology is divided into 5 steps toachieve the desired results:

Step 1: In this step, prepare the data and specify the source of data.

*Step 2:*In this step select the specific data and transform it into different format by weka.

*Step 3:*In this step, implement data mining algorithms and checking of all the relevant bugs and errors is perform.

Step 4: The decision is taken on the presence of bugs insource code. If Bug is present then proceed further,otherwise it will stop.

Step 5: In this step, we make clusters of particular bug orerror with the help of modified K-Means clustering.

Step 6: We classify the relevant bugs using J48graft, Lad tree and Bayes Net algorithm at particular time, after clustering.

Step 7: At the end, the results are display and evaluated.

3.1 Data Preparation

Table1: Represents Attributes of Document Bug for computation

Property	Description	
Source	Name Of A Project Or Department In MASC That Raises The Problem Report.	
Bug Type	(Doc-Bug)The Bug Is From The Software Code Implementation.	
Sample Size	61 Total: 6 Doc-Bug And 55 Non Doc-Bug Software Bug-Tracking System, GNATS (A Tracking System By GNU), Is Set Up On MASC Intranet.	
Dependable Variable	Description	
Severity(1)	Problem Report Is Normal	
Severity(0)	Problem Report Is Serious	
Explanatory Variable	Value	Description
Bug	{1= Bug Accepted,0=Bug Not Accepted}	Describe The Bug Or Defect In The Software
Class	{0=Sw-Bug,1=Doc-Bug,2=Chang Request, 3=Support, 4=Mistaken, 5=Duplicate}	Category Of Bug Class
Priority	{0=Not,1=High,2=Medium,3=Low}	Describe Schedule Permit Duration
Time To Fix	{0=Within Two Days,1=Within One Week,2=Within Two Week,3=Within Three Week,4=Within Four Week,5=Within Five Week}	Take Time Duration In Of Problem Report
State	{0=Closed,1=Open,2=Active,3=Analysed,4=Suspended,5=Resolved,6=Feedback}	Status Of Problem Report Analysis/Non Analysis

A software error arises in problem report and all problem reports can be grouped in two categories: severity and none severity. In severity the data set have no error in software in none severity means a software bug arises which is tracked by GANTS which is a bug tracking system in GNU. It is set on MASC intranet to collect and maintain all problem reports from every department of MASC. The document-bug create in software document categories by class field .Now performing for classification of doc-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 word 2010 database and database management system to store the collect data. 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 step only those fields were selected which were required for data mining. A few derived variables were selected. Where some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from database are given in table 1 for reference. The survey uses status of problem report analysis /non analysis and the operationalization of the survey for items is as follows:

6=Feedback,
5=Resolved,
4=Suspended,
3=Analyzed,
2=Active,
1=Open,
0=Closed

The domain values for some of the variables were defined in the table -1 for the present investigation as follow:

3.3 Data Mining Implementation

Weka is open source software that implements a large collection of machine learning algorithms and is widely used in data mining applications. From the above data bug.arff file was created. This file was loaded into weka explorer and analyzes severity of software defects predicts. Predicts categorical class level classifiers based on training set and the values in the class level attribute use the model in classifying new data. We integrate both (classification and clustering) techniques. After combine application of most frequent used clustering (k-means) algorithm with classification (J48GRAFT, LAD TREE and BAYESNET) algorithms, the results were compare and the weka data mining tool was used.

The problem in particular is a comparative study of performance of integrated clustering and classification technique i.e. Simple k-means clustering algorithm integrated with different classifier such as J48graft,LAD tree and BAYESNET by using various parameters of document- bug, data set containing 5 attributes , 61 instances and one class attribute.

3.4 Result and Discussion-

To better understand the importance of the input variables and analyzed and performance of document –bug .In our research evaluating the performance of above integrated techniques. The data set needs to be normalized by which removing missing values from data set and if any null field present then there will be removed by adding zero instead of null. After normalized the integrate technique applied in which the k-means technique applied by which divide dataset into number of clusters.

```
LogScore MDL: -225.7444329214829
LogScore ENTROPY: -166.1367618909699
LogScore AIC: -195.1367618909699


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          61               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0002
Root mean squared error                  0.0012
Relative absolute error                  0.0972 %
Root relative squared error              0.3963 %
Total Number of Instances               61

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        one
                 1         0          1         1         1          1        zero
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

  a  b   <-- classified as
 55  0 |  a = one
  0  6 |  b = zero
```

Figure 4: J48graft tree for error detection

```
Size of the tree :      7


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          61               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               61

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        one
                 1         0          1         1         1          1        zero
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

  a  b   <-- classified as
 55  0 |  a = one
  0  6 |  b = zero
```

Figure 5: BayesNet for error detection

Classifier output

```
#Leaves (number of predictor nodes): 2
#Expanded nodes: 28
#Processed examples: 1159
#Ratio e/n: 41.392857142857146

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          61               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0.0072 %
Root relative squared error              0.0045 %
Total Number of Instances               61

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        one
                 1         0          1         1         1          1        zero
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

  a  b   <-- classified as
 55  0 |  a = one
  0  6 |  b = zero
```

Figure 6. LAD Tree for error detection

Euclidean distance is a simple distance measure algorithm which calculates the distance of each data values from centroid. Maximum iteration, it is the value that the maximum number of clustering cycle iterates. Number of clusters is what user needs to choose for dividing the data set after clustering the result set needs to be saved in .arff format for applying the classifying algorithm to make integration.

Table 2: K-Mean Error and Time comparison

No. Of clusters	No. Of iterations	Sum of squared errors	Time (seconds)	Data sets
2	2	12	0	19
2	2	27	0	39
2	2	45	0	61

From the above Table 3 it may be observed that k-means+J48GRAFT and k-means + BAYESNET both take less time than K-Means + Lad tree. However K-Means + BayesNet has 0.0 MAE, and 0.0 time taken by k-means+J48GRAFT has 0.0 MAE and 0.02 time taken

Table 3: Represents Error Comparison with K-Means between J48graft, Lad tree and Bayes Net

Detailes	K-Means+J48graft	K-Means+Lad Tree	K-Mean+Bayesnet
Mean Absolute Error	0	0.0	.0002
Root Mean Square Error	0	0.0	0.0012
Time Taken	0.02	0.03	0

According to test and outcome analysis in our research, it was found that in the integrated approach of clustering and classification, the performance of K-Means + BayesNet is better than other algorithms on the basis of time but on the basis of error k-means perform better than other algorithms.

4. CONCLUSION

In this paper, three different classifier are integrated with the simple k-means clustering algorithm and integration techniques were applied on document bug data set. From the observation and analysis it was concluded that the integration of K-Means + BayesNet have 0.0002 MAE and 0.0012 RMSE error and it also takes 0.0 less time to build the model. So on the basis of time and error we found K-Means + BayesNet is better than other. There is large number of classifiers presents and data mining cluster are present. So the future work will be based on other classifier than can be applied on the real data set and also to apply other data mining tools on the data set such that the best techniques can be identified.

REFERENCES

1. Shi, Haijian, “Best-first Decision Tree Learning”, Master’s Degree Theses. University of Waikato Masters Theses, 2007.
2. D. Jachayra, K .Pancerz and J. Gomula, “Classification of MMPI profile using decision tree”, Concurrency specification and programming, 2011.
3. D. Buhmann, “Radial Basis function: Theory and Implementation”, Cambridge monographs on applied and computational mathematics,2003.
4. G.J.Pai, J.B.Dugan, “Empirical Analysis of Software Fault Content and Fault Proneness Using Bayesian Methods”, IEEE Trans. Software Eng., vol. 33, no.10, pp. 675- 686, July2007.
5. Sunita Tiwari and Neha Chaudhary “Data mining And Warehousing”, Dhanpati Rai and Co.(P) Ltd. First edition: 2010.
6. M. Shepperd, C. Schofield, and B. Kitchenham, “Effort estimation using analogy”, in of the 18th International Conference on Software Engineering, pp.170-178. Berlin, Germany, 1996.
7. Yadav, Surjeet Kumar, and Saurabh Pal. "Data mining application in enrollment management: A case study." International Journal of Computer Applications (IJCA) 41.5, 2012, 1-6.
8. Pal, A. K., & Pal, S., “Classification model of prediction for placement of students”, International Journal of Modern Education and Computer Science (IJMECS), 5(11),2013, 49.
9. Alsmadi and Magel, “Open source evolution Analysis,” in proceeding of the 22nd IEEE International Conference on Software Maintenance (ICMS’06), Philadelphia, pa.USA, 2006.
10. Boehm, Clark, Horowitz, Madachy, Shelby and 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.Artherand S.M. Henry, Eds, vol.1, pp.45-60,j.c. Baltzer AG, science publishers, Amsterdam, The Netherlands, 1995.

11. Ribu,Estimating, “Object oriented software projects with use cases”, M.S. thesis, University of Oslo Department of informatics, 2001.
12. N.Nagwani and S. Verma, “Prediction data mining model for software bug estimation using average weighted similiarity,” In proceeding of advance computing conference(IACC),2010.
13. A.E.Hassan, “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.
14. Yadav, Dhyan Chandra, and Saurabh Pal. "Analysis Receiver Operating Characteristics of Software Quality Requirement by Classification Algorithms Analysis” IJCA 116.8, 2015.
15. Yadav, Dhyan Chandra, and Saurabh Pal. "Software Bug Detection using Data Mining." IJCA, 115.15, 2015
16. Z.Li and Reformat, “A practical method for the Software fault prediction”, in proceeding of IEEE Nation conference information reuse and Integration (IRI), 2007.
17. C. Elcan, “The foundations of cost sensitive learning”, in processing of the 17 International conferences on Machine learning, 2001.
18. C.Chang and C.Chu, “software defect prediction using international association rule mining”, 2009.
19. S. Kotsiantis and D. Kanellopoulos, “Association rule mining: A recent overview”, GESTS international transaction on computer science and engineering, 2006.
20. N. Pannurat, N. Kerdprasop and K. Kerdprasop, “Database reverses engineering based on Association rule mining”, IJCSI international journal Of computer science issues 2010.
21. Fayyad, Piatesky Shapiro, Smuth and Uthurusamy, “Advances in knowledge discovery and data mining”, AAAI Press,1996.
22. M.Shtern and Vassilios, “Review article advances in software engineering clustering methodologies for software engineering”, Tzerpos volume, 2012.
23. P.Runeson and O.Nyholm, “Detection of duplicate defect report using neural network processing”, in proceeding of the 29th international conference on software engineering 2007.
24. G.Vishal and S.L. Gurpreet, “A servey of text mining techniques and applications”, journal of engineering technologies in web intelligence, 2009.
25. Lovedeep and Varinder Kaur Arti, “Application of data mining techniques in software engineering” International journal of electrical, electronics and computer system(IJEECS) volume-2 issue-5,6. 2014.
26. Richi Nayak and Tian Qiu, “A data mining application,” international journal of software engineering Knowledge engineering, volume.15, issue-04,2005.
27. 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.
28. Chauraisa V. and Pal S., “Early Prediction of Heart Diseases Using Data Mining Techniques”, Carib.j.SciTech,,Vol.1, pp. 208-217, 2013.

Vol 8. No. 2 June, 2015

African Journal of Computing & ICT

www.ajocict.net

An Integration of Clustering and Classification Technique in Software Error Detection

D.C. Yadav
Research Scholar,
Shri Venkatwhwara University, Amroha
E-mail: dc9532105114@gmail.com

Pal, S.
Department of MCA,
VBS Purvanchal University Jaunpur (U.P.)
E-mail: drsaurabhpal@yahoo.co.in

ABSTRACT

The software project development plays important role in software quality. Measuring software quality in a specific manner such as error estimation and check severity of error which is related to the document bug .Data mining create a supportable platform for software project development by which software engineers easily achieved goal of project quality in given time duration and budget .In this paper we integrate both technique classification and clustering for software error detection. Classification technique analyzes the severity of software defect by J48GRAFT, LAD Tree, and BAYESNET also by Clustering technique measure maximum similar data object in data set within same cluster by K-Means.

Keywords - Data mining; Classification: J48graft, Lad Tree and BayesNet; Clustering: K-Means; Weka.

African Journal of Computing & ICT Reference Format:
D.C. Yadav & S. Pal (2015). An Integration of Clustering and Classification Technique In Software Error Detection.
Afr J. of Comp & ICTs. Vol 8, No. 2. Pp9-16.

I. INTRODUCTION

Software testing check the performance of software project in which, software tracker plays a technical role to detect the software defect. The software tracker provides feedback information about bug. And give the technical role in software quality improvement. Software tracker provides the severity of bug fix or not fix in the software. Tracker measuring software quality in a specific manner such as error estimation and check the severity of error which is related to the document is known as document bug also decide the severity of bug. Shi and Harjan [1] in 2007 presented data classification of software defect is very commanding mission in data mining. The categories of software defects in data mining used comparable tree function, rule etc. The purpose of arrangement is to analyze the ideas of each adjustable and allocate those variables to corresponding predefined classes.

Jachyra, Pancerz and Gomula [2] introduced about J48graft. J48graft generates a technical graphical way to describe unrecovered problems of document bug in training set. The J48graft generates decision tree and give the training set but it have some drawback which is recover by J48graft. J48graft classify the environment into multi-dimensional space which is not possible by training set it is also reduce the prediction error. In the Fig. 1 J48graft classify the bug of software defect and easily we analyzed bug by J48graft tree.