Volume 1, Issue 1, January 2014 ISSN: 2348-8565 (Online)

International Journal of Modern Engineering and Research Technology

Website: http://www.ijmert.org Email: editor.ijmert@gmail.com

Analysis of Different Clustering Algorithms on Different Software Defect

Dhyan Chandra Yadav
Research scholar,
Shri Venkateshwara University
Gajraula, Amroha (U.P.), India
Email: dc9532105114@gmail.com

Rajeev Kumar
Assistant Professor,
Department of Computer Science.
Shri Venkateshwara University
Gajraula, Amroha (U.P.), India
E-mail: rajeev2009mca@gmail.com

ABSTRACT

Data clustering is not a single things but it is a process of positioning similar data into groups. A gathering algorithm partitions a data set interested in several collections based on the attitude of take full advantage of the intra-class similarity and reducing the inter-class similarity. This paper analyze the three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compare the performance of these three major clustering algorithms on the software defect data set and measure cluster building ability of algorithm. Performance of these techniques are presented and compared using a clustering tool WEKA.

Keywords:— *Clustering Algorithms: K-means algorithms, Hierarchical clustering, Density based clustering algorithm, Weka.*

I. INTRODUCTION

Anthony Williams [1] Repetition is a technical issue in the life cycle of a software project development. Duplicate is closed as many spaces between codes, change code elsewhere and error at the bottom of the line All technical problems reported in a report known as problems report .If any bug is reported in the problem report but is already covered by another problem report this occurrence is known as a repetition error. A duplicate bug is created in any phase test and is sometimes automatically generated for coding or phase testing with the help of data mining. It is easily classified and analyzed in the software engineering field.

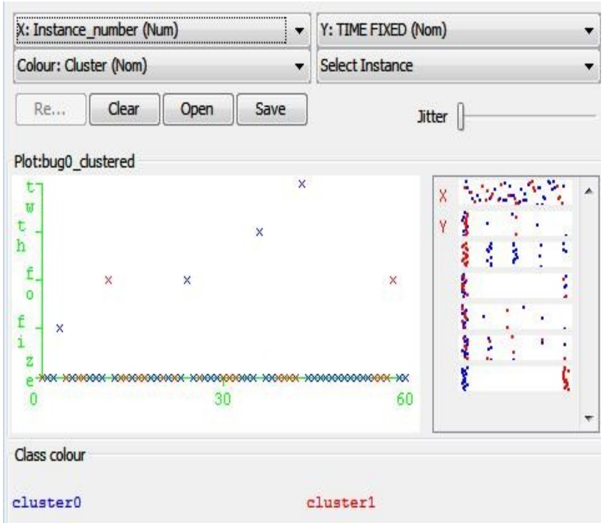
1.1. Clustering Algorithms

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.

Many algorithms exist for clustering. Following figures showing three major clustering methods and their approach for clustering.

1.1.1 K-means Clustering

MacQueen J. B. [2] is introduced when the mahattam range is used and the centroid is calculated as part of medium smarter than methods. K-means is a widely used method of classification. K-means algorithm is the most widely used segmentation algorithm because it can be easily implemented and is the most efficient in terms of execution. For example we visualize featured data set as:



X: Instance_number (Num)
Y: TIME FIXED (Nom)
Colour: Cluster (Nom)
Select Instance

Figure 1. Visualize of K-means Clustering

1.1.2 Hierarchical clustering

Mansh Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta [3] presented about Sequential collection using the classic value at the top. Hierarchical Clustering forms a cluster or, in other words, a cluster tree. An example explains:

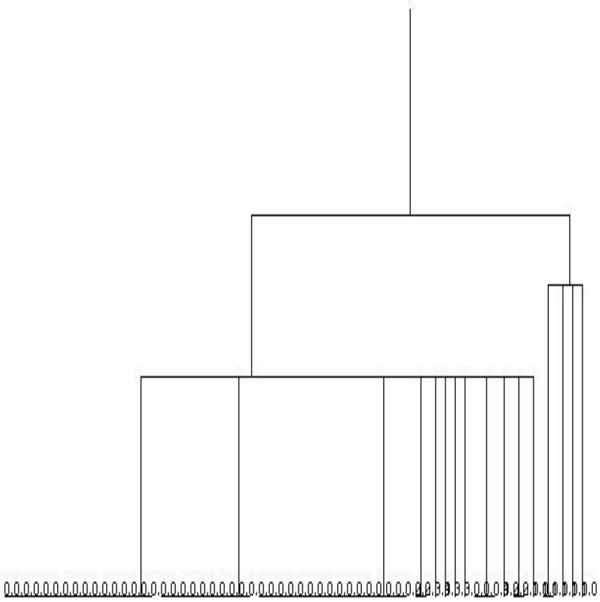


Figure 2. Visualize of Hierarchical clustering

- ***Agglomerative (bottom up)***

1. It starts with a singleton point.
2. Repeat to add two or more suitable groups.
3. Stop where the number of k groups is found.

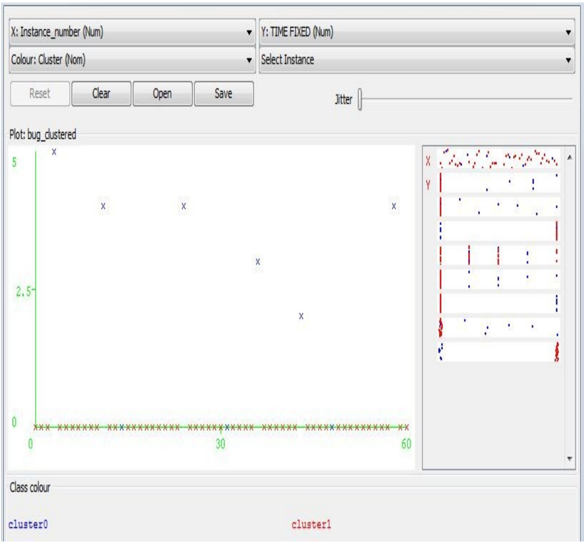
- ***Divisive (top down)***

1. Start with a large collection.
2. It is often divided into smaller groups.
3. Stop where the number of k groups is found.

This figure show that the result of Hierarchical clustering Methods with single-linkage [7] between data points using WEKA tool.

- ***Density based clustering***

Timonthy C. Havens [4] presented about. It only supports the number of clusters the interface is visible when the integrated cluster does. Strength collection algorithms attempt to obtain collections based on increasing data facts in a section. The basic idea of mass compilation is that in each case the collection area should be at least containing a minimum number of cases (MinPts).



X: Instance_number (Num)
Y: TIME FIXED (Num)
Colour: Cluster (Nom)
Select Instance

Figure 3. Visualized of Density based Clustering.

Above figure showing the result of density-based clustering methods using WEKA tool.

II. Related Work

Shepperd, Schofield and Kitchenham [5] discussed the need to estimate the cost of management and software development organizations and provided the concept of forecasting and rating methods.

Alsmadi and Magel [6] discussed how data mining provides space for a new software project with its quality, cost and complexity also forming a channel between data mining and software engineering.

Boehm, Clark, Horowitz, Madachy, Shelby and Westland [7] discussed that some software companies are experiencing some accuracy issues depending on his data set after the forecasting software company offered a new idea to define a project cost schedule and determine a staff time table.

K.Ribu [8] discussed the need for open source code projects that are analyzed by predicting and discovering a software project that focuses on the object with case models.

Nagwani and Verma [9] discussed that the software bug (bug) prediction and bug and bug rating for all software summaries, with data mining also discussed the software bug.

Hassan [10] argued that a complex data source (audio, video, text, etc.) requires a lot of buffer to process and does not support standard size and buffer length.

Li and Reformate [11] discussed that .the software configuration management system includes documents, software code, case statistics, tracking and incorporating update data.

Elcan [12] argued that the COCOMO model was undermined by accurate cost estimates and there is a lot about cost estimation because in project development it involves a lot of flexibility and therefore a COCOMO measurement over time and metrics.

Chang and Chu [13] discussed that in terms of finding a large database pattern and its variability and the relationship between them in the organization of data mining organization.

Kotsiantis and Kanellopoulos [14] discussed this major flaw in the development of a software project and also discussed a pattern to provide space for prediction and integrated governance to reduce the pass rate in the database.

Pannurat, N. Kerdprasop and K. Kerdprasop [15] discussed that organizational law provides for the relationship between large databases such as software project time, cost record and assistance in project implementation.

Fayyad, Piatesky Shapiro, Smuth and Uthurusamy [16] discussed that separation creates a relationship or map between the data object and the classes described earlier.

Stern and Vassillios [17] argued that in the analysis of the combination of the same object placed in the same group it also classified the attribute in the group so that the differences between the groups were increased in relation to the differences within the groups.

Runeson and Nyholm [18] have argued that code replication is an independent language problem. Repeatedly another report of a problem in software development and duplication arises using a neural language with a data mine.

Vishal and Gurpreet [19] argued that the data mine analyzes data and researches hidden data from the text in the development of a software project.

The Lovedeep and Arti data mine [20] offers a specific software engineering platform where most functions are easily operated with the best quality and reduce the cost and problems of high profile.

Nayak and Qiu [21] discussed that often time and cost, related problems arise from software project development these problems identified in the problem report, data mining provides help in reducing problems and isolating and reducing other software-related bugs.

This paper analyzes three major integration algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compares the performance of these three major integration systems in a software data setting and measures the ability to create a set of algorithms. The implementation of these strategies was presented and compared using the WEKA integration tool.

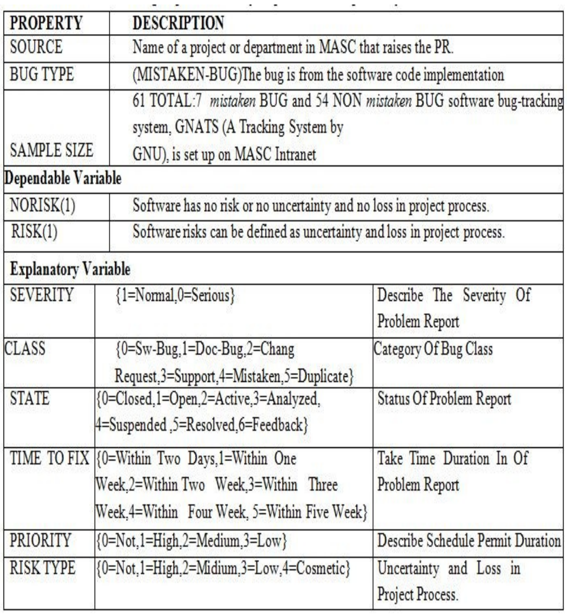
III. Methodology

The data sets are collected and processed after which it is converted into a suitable format. Post this step, clustering is done and results are displayed in an understandable format. Our application helps to bridge the gap between the consumer and the software bug by giving the software information on consumers' choices and preferences.

3.1. Data Preparation

A software error occurs in the problem report and all problem reports collected in two categories: available and unavailable. In the recoverable group the error was easily detected automatically with the software bug tracking system GANTS. Set up on the MASC intranet to collect and store all problem reports from all MASC departments.

Table 1: Mistaken Bug Representation by Dependable & Explanatory Variables

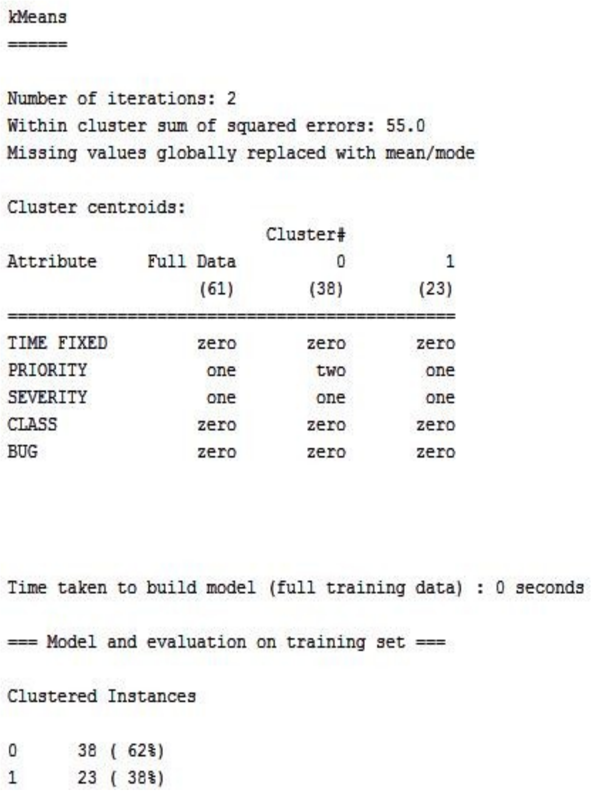


PROPERTY	DESCRIPTION	
SOURCE	Name of a project or department in MASC that raises the PR.	
BUG TYPE	(MISTAKEN-BUG)The bug is from the software code implementation	
SAMPLE SIZE	61 TOTAL:7 *mistaken* BUG and 54 NON *mistaken* BUG software bug-tracking system, GNATS (A Tracking System by GNU), is set up on MASC Intranet	
Dependable Variable		
NORISK(1)	Software has no risk or no uncertainty and no loss in project process.	
RISK(1)	Software risks can be defined as uncertainty and loss in project process.	
Explanatory Variable		
SEVERITY	{1=Normal,0=Serious}	Describe The Severity Of Problem Report
CLASS	{0=Sw-Bug,1=Doc-Bug,2=Chang Request,3=Support,4=Mistaken,5=Duplicate}	Category Of Bug Class
STATE	{0=Closed,1=Open,2=Active,3=Analyzed, 4=Suspended ,5=Resolved,6=Feedback}	Status Of Problem Report
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
PRIORITY	{0=Not,1=High,2=Medium,3=Low}	Describe Schedule Permit Duration
RISK TYPE	{0=Not,1=High,2=Midium,3=Low,4=Cosmetic}	Uncertainty and Loss in Project Process.

If a bug is already covered by another report of a problem it is known as a duplicate-bug. Duplicate-bug arises from the use of code. Duplicate-bug classification is now done using a number of standard data mining operations, data processing, integration, classification, merging and operations need to be done. The database is built into the MS-Excel database, MS word 2010. The data is compiled according to the required format and formats and the data is converted to ARFF (affiliate file format) for processing to set. The ARFF file is an ASCII text file that describes a list of situations that share a set of symbols.

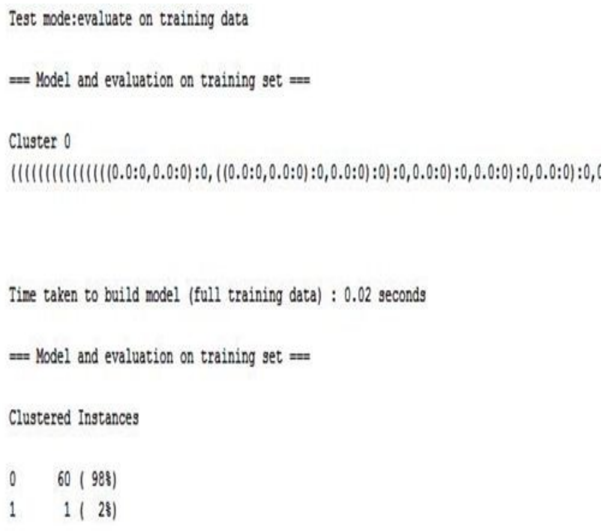
3.2. Data Selection and Transform

Often finding the best learning program for a given task is a matter of trial and error. Several techniques will need to be tested with different parameters, and their results have been analyzed to find the most suitable. Experimenter is used to make this process automated, it can queue up multiple machine learning algorithms, to run on multiple data sets and collect statistics on their performance. By performing a collection analysis on weka. I uploaded a set of data set to the image shown in the image. By default the data set must have it in CSV format.



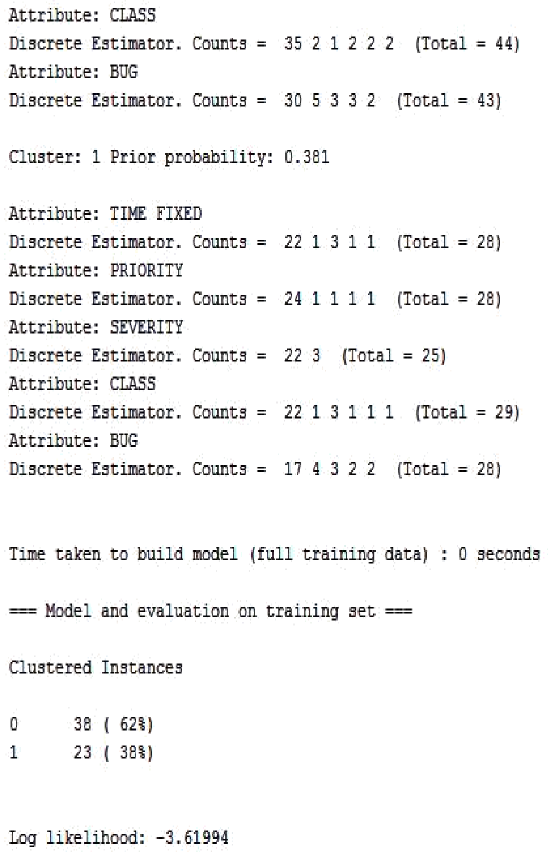
```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 55.0
Missing values globally replaced with mean/mode

Cluster centroids:
                              Cluster#
Attribute     Full Data          0           1
                   (61)       (38)        (23)
=============================================
TIME FIXED         zero       zero        zero
PRIORITY            one        two         one
SEVERITY            one        one         one
CLASS              zero       zero        zero
BUG                zero       zero        zero




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      38 ( 62%)
1      23 ( 38%)
```

Figure 4: Representation of Simple K-Means clustering



```
Test mode:evaluate on training data

=== Model and evaluation on training set ===

Cluster 0
(((((((((((((0.0:0,0.0:0):0,((0.0:0,0.0:0):0,0.0:0):0):0,0.0:0):0,0.0:0):0,0.0:0):0,(



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      60 ( 98%)
1       1 (  2%)
```

Figure 5: Representation of Hierarchical Clustering



```
Attribute: CLASS
Discrete Estimator. Counts =  35 2 1 2 2 2  (Total = 44)
Attribute: BUG
Discrete Estimator. Counts =  30 5 3 3 2  (Total = 43)

Cluster: 1 Prior probability: 0.381

Attribute: TIME FIXED
Discrete Estimator. Counts =  22 1 3 1 1  (Total = 28)
Attribute: PRIORITY
Discrete Estimator. Counts =  24 1 1 1 1  (Total = 28)
Attribute: SEVERITY
Discrete Estimator. Counts =  22 3  (Total = 25)
Attribute: CLASS
Discrete Estimator. Counts =  22 1 3 1 1 1  (Total = 29)
Attribute: BUG
Discrete Estimator. Counts =  17 4 3 2 2  (Total = 28)


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      38 ( 62%)
1      23 ( 38%)


Log likelihood: -3.61994
```

Figure 6: Representation of DBSCAN clustering

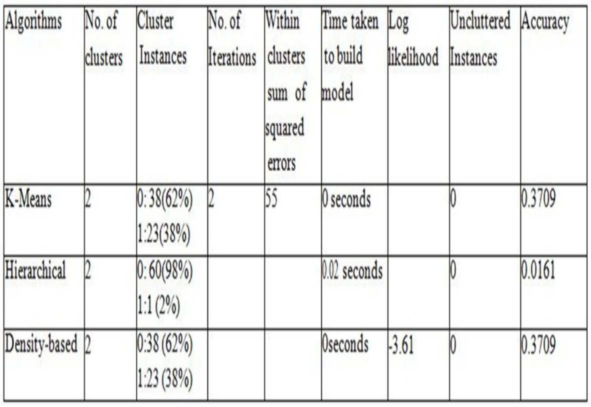
3.3. Data Mining Implementation

Weka is the data mining tool to classify data into different types. It is the first model to provide a graphical user interface. To make the merger we use the promise data area. Provides details of a previous analysis project. With the help of statistics we show the effectiveness of the various algorithms used in weka. We set the most suitable tool for data mining applications. This paper only shows weka compilation functions, we will try to make a complete weka reference paper. Assembling is a major metaphor for mining data, as well as the standard mathematical data analysis method used in many fields, including machine learning. I use Enter data input tools for this purpose. It provides a batter interface to the user rather than comparing other data mining tools.

3.4. Results and discussion-

The section above involves a study of each of the three previously introduced methods using the Collection Tool in a bank data set with 9 symbols and 61 entries. Data set integration is done with each merging algorithm using the Keep tool and the results are:

Table 2: Comparison result of algorithms using weka tool



Algorithms	No. of clusters	Cluster Instances	No. of Iterations	Within clusters sum of squared errors	Time taken to build model	Log likelihood	Uncluttered Instances	Accuracy
K-Means	2	0: 38(62%) 1:23(38%)	2	55	0 seconds		0	0.3709
Hierarchical	2	0: 60(98%) 1:1 (2%)			0.02 seconds		0	0.0161
Density-based	2	0:38 (62%) 1:23 (38%)			0seconds	-3.61	0	0.3709

From the table above it is clear that K-Means and the making of moderately based combinations provide more accurate results compared to Hierarchical integration. K-Means offers 2 sets and 2 repetitions without a Log Likelihood. Make a compact-based collection with the negative value of Log Likelihood which shows fewer opportunities. K-Means does not offer bad Log Likelihood prices.

V. CONCLUSION

After analyzing the test results of the algorithms we can draw the following conclusions: The performance of the K-Means algorithm is better than the Hierarchical Clustering and Make density based algorithm. All algorithms have ambiguity in some (bug) data when combined. The robustness algorithm is not suitable for data with high density variability.

REFERENCES:

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

[2] MacQueen J. B., "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. 1967, pp. 281–297.

[3] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.

[4] Timonthy C. Havens. "Clustering in relational data and ontologies" July 2010.

[5] 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.

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

[7] 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. Arther and S.M. Henry, Eds, vol.1, pp.45-60, J.C. Baltzer AG, science publishers, Amsterdam, The Netherlands, 1995.

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

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

[10] 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.

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

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

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

[14] S. Kotsiantis and D. Kanellopoulos, "Associan rule mining: Arecentover view", GESTS international transaction a on computer science and Engineering, 2006.

[15] N.Pannurat, N.Kerdprasop and K.Kerdprasop "Database reverses engineering based On Association rule mining", IJCSI international Journal of computer science issues 2010.

[16] U.M.Fayyad, G. Piatesky Shapiro, P.Smuth and R.Uthurusamy, "Advances in knowledge discovery And data mining", AAAI Press,1996.

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

[18] 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.

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

[20] 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.

[21] Richi Nayak and TianQiu "Adata mining application" international journal of software Engineering and Knowledge engineering volume.15, issue-04, 2005.

* * * * *