Enhanced Bug Detection by Data Mining Techniques

Promila Devi[1], Rajiv Ranjan*[2]
*[1] M.Tech(CSE) Student, *[2] ASSISTANT PROFESSOR(CSE) Arni University, Indora, Kangra, India

ABSTRACT

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data Mining in Java is a big challenging problem nowadays, When an application opens which consist of large data in database it takes so much time to load and get that data however it may contain large number of bugs. So the problem with " Big Data Mining " is still a issue. We have to rectify this issue with effective approach with decision tree classifier in which we need clustering of data with k-means error and bug search of that particular source code of application. We will enhance the search on Bug Detection the K-means clustering algorithm with the help of multi-threading Decision Tree. In this work of research the problem with classification of bugs is been identified.

KEYWORDS: *Data Mining, Clustering Analysis, Partitioning, Clustering, K-means, decision tree.*

I. INTRODUCTION

Data mining techniques have increasingly been studied[7] especially in their application in real-world databases. One typical problem is that databases tend to be very large, and these techniques often repeatedly scan the entire set. Sampling has been used for a long time, but subtle di_erences among sets of objects become less evident.This work provides an overview of some important data mining techniques and their applicability on large databases.

Objectives

[1] To maintain the sustainability of the code of object oriented languages.
[2] To clearly help the testers to classify the number of bugs present in a source code.
[3] Increasing the response/execution time of proposed algorithm using decision tree.
[4] To attain the accuracy which helps in future growth of testing phase of various IT applications and smart phone applications.

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

C Programming Error types: While writing c programs, errors also known as bugs in the world of programming may occur unwillingly which may prevent the program to compile and run correctly as per the expectation of the programmer.Basically there are three types of errors in c programming:

[1] Runtime Errors
[2] Compile Errors
[3] Logical Errors

C Runtime Errors: C runtime errors are those errors that occur during the execution of a c program and generally occur due to some illegal operation performed in the program.Examples of some illegal operations that may produce runtime errors are***:***

- Dividing a number by zero
- Trying to open a file which is not created
- Lack of free memory space

It should be noted that occurrence of these errors may stop program execution, thus to encounter this, a program should be written such that it is able to handle such unexpected errors and rather than terminating unexpectedly, it should be able to continue operating. This ability of the program is known as **robustness** and the code used to make a program robust is known as **guard code** as it guards program from terminating abruptly due to occurrence of execution errors.

Compile Errors:Compile errors are those errors that occur at the time of compilation of the program. C compile errors may be further classified as:

Syntax Errors:When the rules of the c programming language are not followed, the compiler will show syntax errors.

For example, consider the statement,

```
1 int a,b:
```

The above statement will produce syntax error as the statement is terminated with : rather than ;

Semantic Errors: Semantic errors are reported by the compiler when the statements written in the c program are not meaningful to the compiler.

For example, consider the statement,

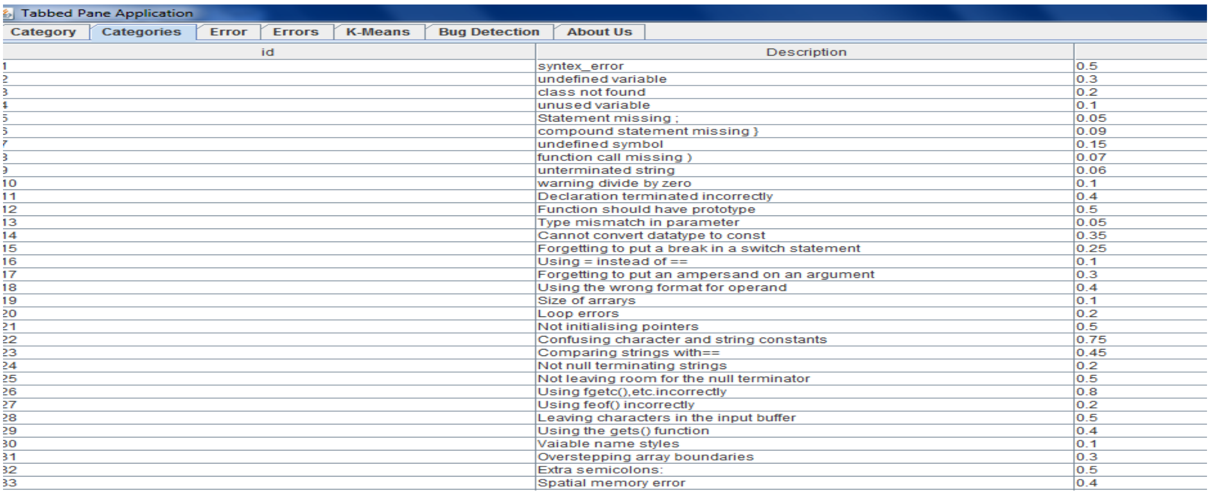
```
1  b+c=a;
```

In the above statement we are trying to assign value of a in the value obtained by summation of b and c which has no meaning in c. The correct statement will be

```
1  a=b+c;
```

Logical Errors:Logical errors are the errors in the output of the program. The presence of logical errors leads to undesired or incorrect output and are caused due to error in the logic applied in the program to produce the desired output.

Entries in tab pan:Ther are some error category.

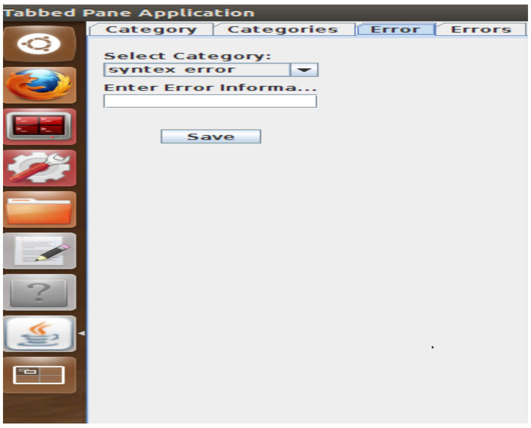


id	Description	
1	syntex_error	0.5
2	undefined variable	0.3
3	class not found	0.2
4	unused variable	0.1
5	Statement missing ;	0.05
6	compound statement missing }	0.09
7	undefined symbol	0.15
8	function call missing)	0.07
9	unterminated string	0.06
10	warning divide by zero	0.1
11	Declaration terminated incorrectly	0.4
12	Function should have prototype	0.5
13	Type mismatch in parameter	0.05
14	Cannot convert datatype to const	0.35
15	Forgetting to put a break in a switch statement	0.25
16	Using = instead of ==	0.1
17	Forgetting to put an ampersand on an argument	0.3
18	Using the wrong format for operand	0.4
19	Size of arrarys	0.1
20	Loop errors	0.2
21	Not initialising pointers	0.5
22	Confusing character and string constants	0.75
23	Comparing strings with==	0.45
24	Not null terminating strings	0.2
25	Not leaving room for the null terminator	0.5
26	Using fgetc(),etc.incorrectly	0.8
27	Using feof() incorrectly	0.2
28	Leaving characters in the input buffer	0.5
29	Using the gets() function	0.4
30	Vaiable name styles	0.1
31	Overstepping array boundaries	0.3
32	Extra semicolons:	0.5
33	Spatial memory error	0.4

•

There are some snap shots:

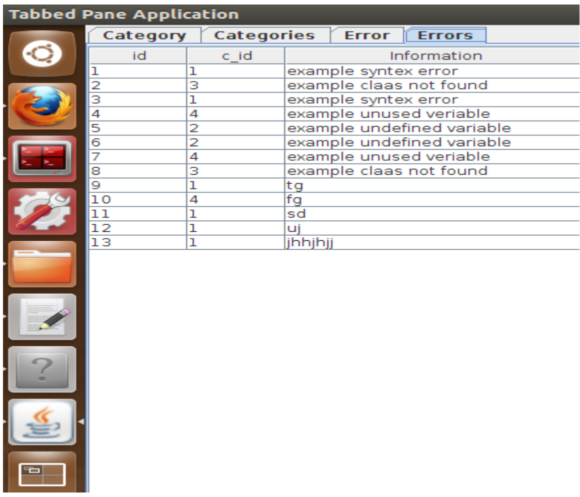
(a) Error Tabbed: First of all we enter name of category and severity and save this for next process.



Select Category:
syntex error
Enter Error Informa...

Fig 1.1

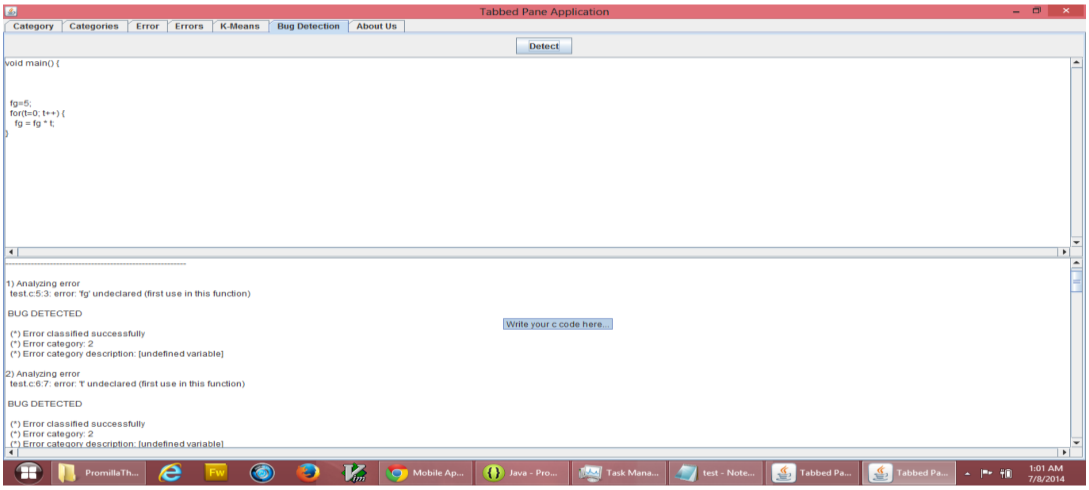
(b) Errors Tabbed: In this tab we enter error category and description of error for example semi colon is missin,error in line 2etc.



id	c_id	Information
1	1	example syntex error
2	3	example claas not found
3	1	example syntex error
4	4	example unused veriable
5	2	example undefined variable
6	2	example undefined variable
7	4	example unused veriable
8	3	example claas not found
9	1	tg
10	4	fg
11	1	sd
12	1	uj
13	1	jhhjhjj

Fig1.2

Bug Detection: Firstly we show the snapshot which represents the Bug detection by using K-means clustering. In this snap shot we detect bug by using bug detection tab and write our program in C Language in code window after writing our code we will use detect button then it will show error if is there any error in our program in the output window **.** In the code window it will show output or detect four errors and display total number of errors.

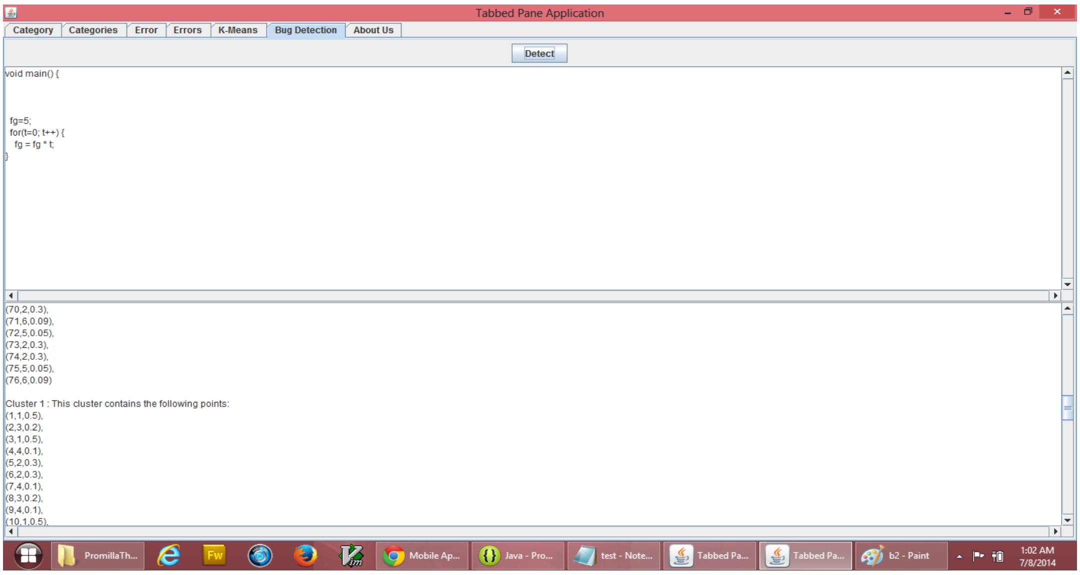


```
void main() {



 fg=5;
 for(t=0; t++) {
   fg = fg * t;
}
```

```
--------------------------------------------------------------

1) Analyzing error
 test.c:5:3: error: 'fg' undeclared (first use in this function)

 BUG DETECTED

 (*) Error classified successfully
 (*) Error category: 2
 (*) Error category description: [undefined variable]

2) Analyzing error
 test.c:6:7: error: 't' undeclared (first use in this function)

 BUG DETECTED

 (*) Error classified successfully
 (*) Error category: 2
 (*) Error category description: [undefined variable]
```

Fig 1.3

Now In this code window K-means cluster are made by above code values.

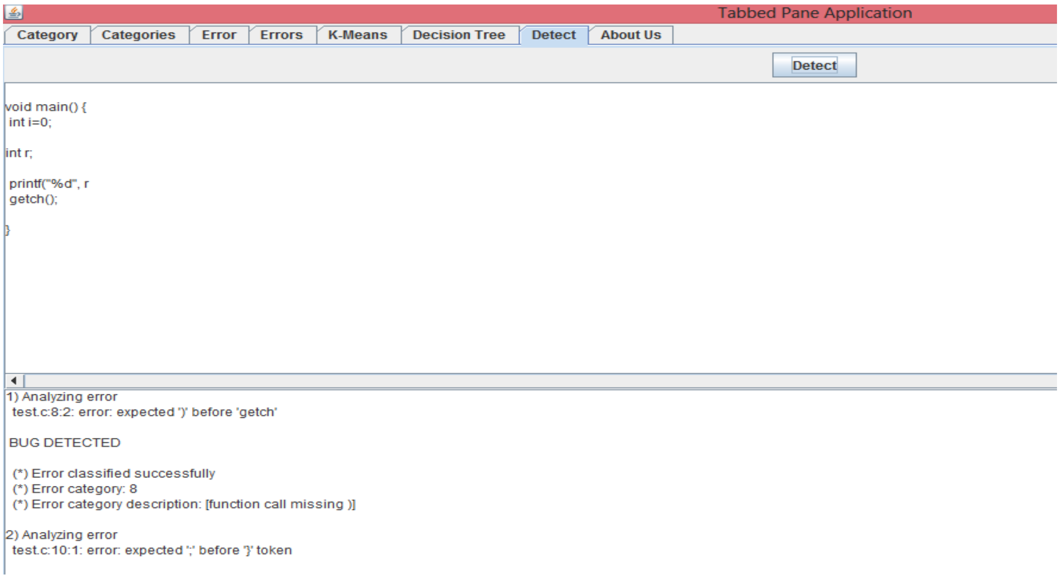


```
void main() {



 fg=5;
 for(t=0; t++) {
   fg = fg * t;
}
```

```
(70,2,0.3),
(71,6,0.09),
(72,5,0.05),
(73,2,0.3),
(74,2,0.3),
(75,5,0.05),
(76,6,0.09)

Cluster 1 : This cluster contains the following points:
(1,1,0.5),
(2,3,0.2),
(3,1,0.5),
(4,4,0.1),
(5,2,0.3),
(6,2,0.3),
(7,4,0.1),
(8,3,0.2),
(9,4,0.1),
(10,1,0.5),
```

Fig 1.4

Now In this code window it will disply predicting the error probability using decision tree & K-Means result

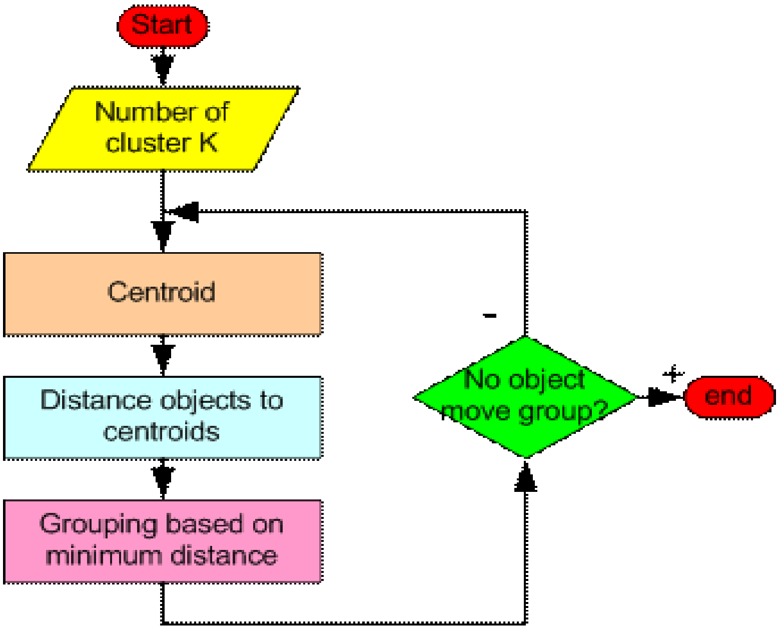


```
void main() {
 int i=0;

int r;

 printf("%d", r
 getch();

}
```

```
1) Analyzing error
 test.c:8:2: error: expected ')' before 'getch'

 BUG DETECTED

  (*) Error classified successfully
  (*) Error category: 8
  (*) Error category description: [function call missing )]

2) Analyzing error
 test.c:10:1: error: expected ';' before '}' token
```

Fig 1.5

How the K-Mean Clustering algorithm works:



Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Fig 1.6

K-means Clustering:

Complexity is O(n * K * I * d)

– n = number of points, K = number of clusters,
I = number of iterations, d = number of attributes

– Easily parallelized

– Use kd-trees or other efficient spatial data structures for some situations

□ Pelleg and Moore (X-means)

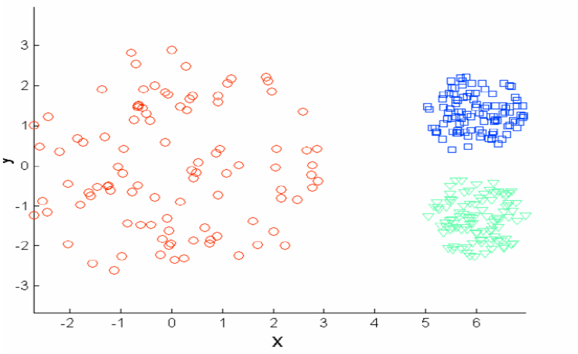
□ Sensitivity to initial conditions

Limitations of K-means:

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes
- Problems with outliers
- Empty clusters

Limitations of K-means: Differing Density

Original Points:



K-means (3 Clusters):

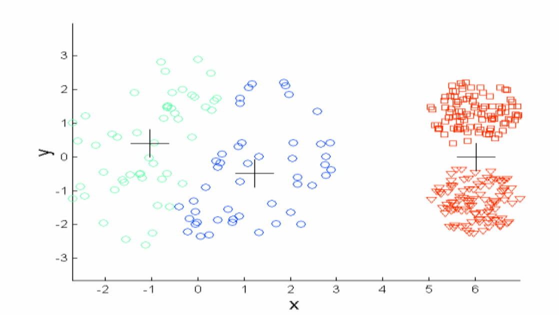
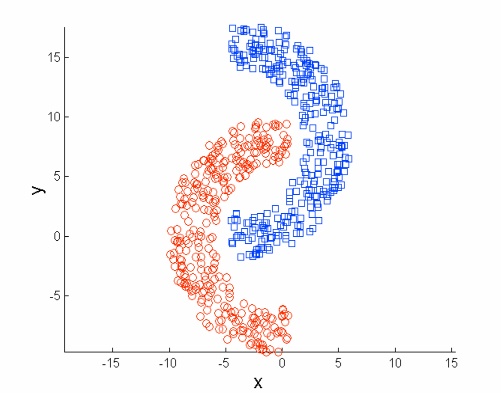


Fig 1.7

Limitations of K-means: Non-globular Shapes:

Original Points:



K-means (2 Clusters):

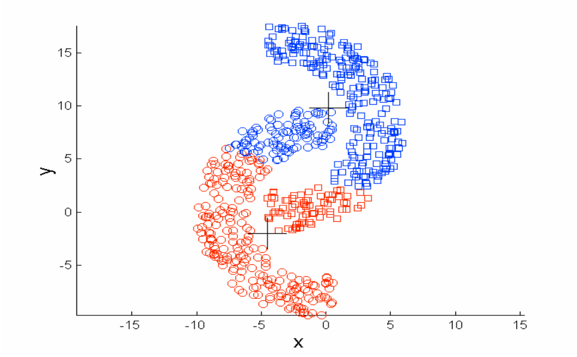
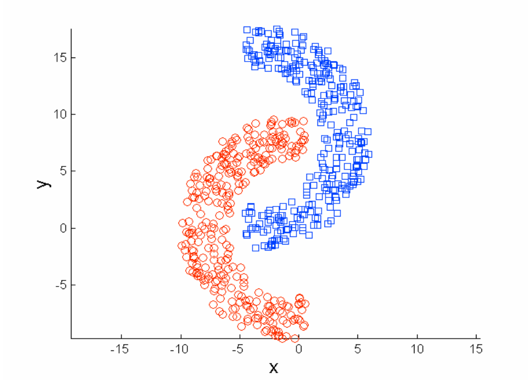


Fig1.8

Overcoming K-means Limitations:

Original Points:



K-means Clusters:

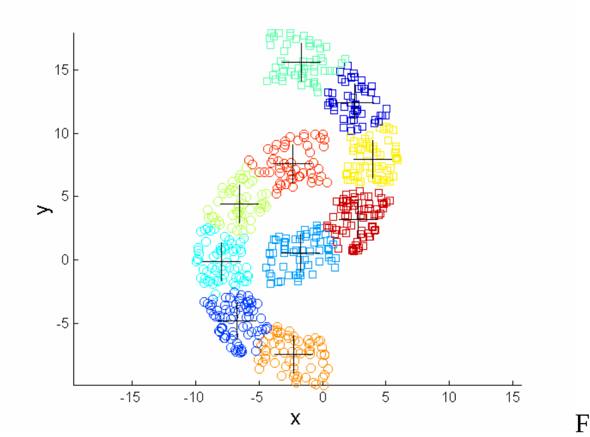
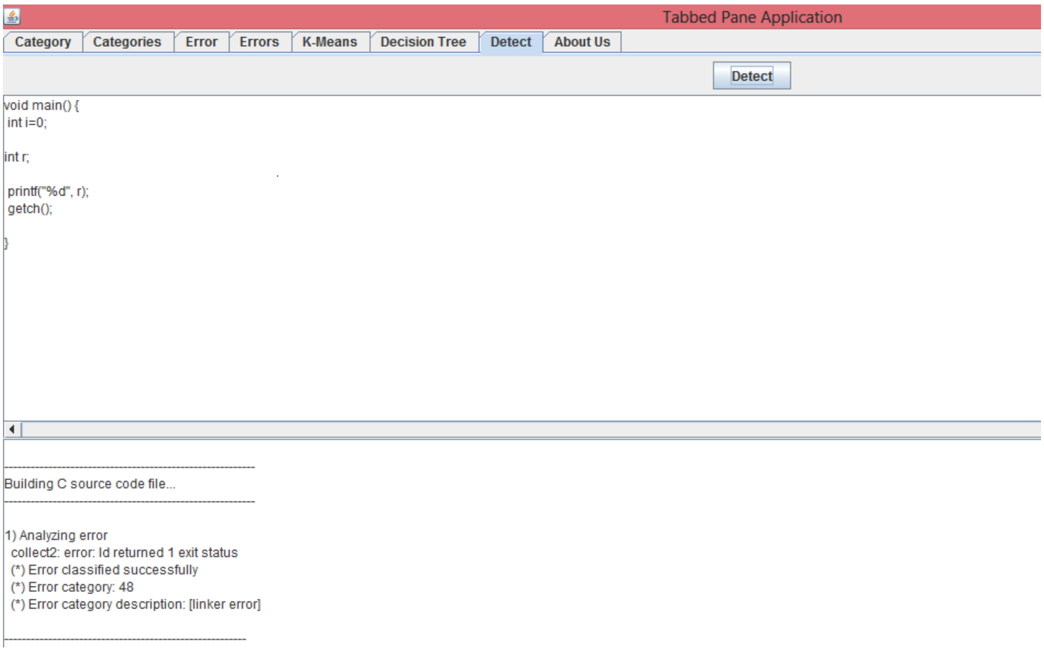


Fig 1.9

Clustering Analysis

Clustering is the division of data into groups containing similar objects. It is used in fields such as pattern recognition, and machine learning [2]. Searching for clusters involves unsupervised learning.

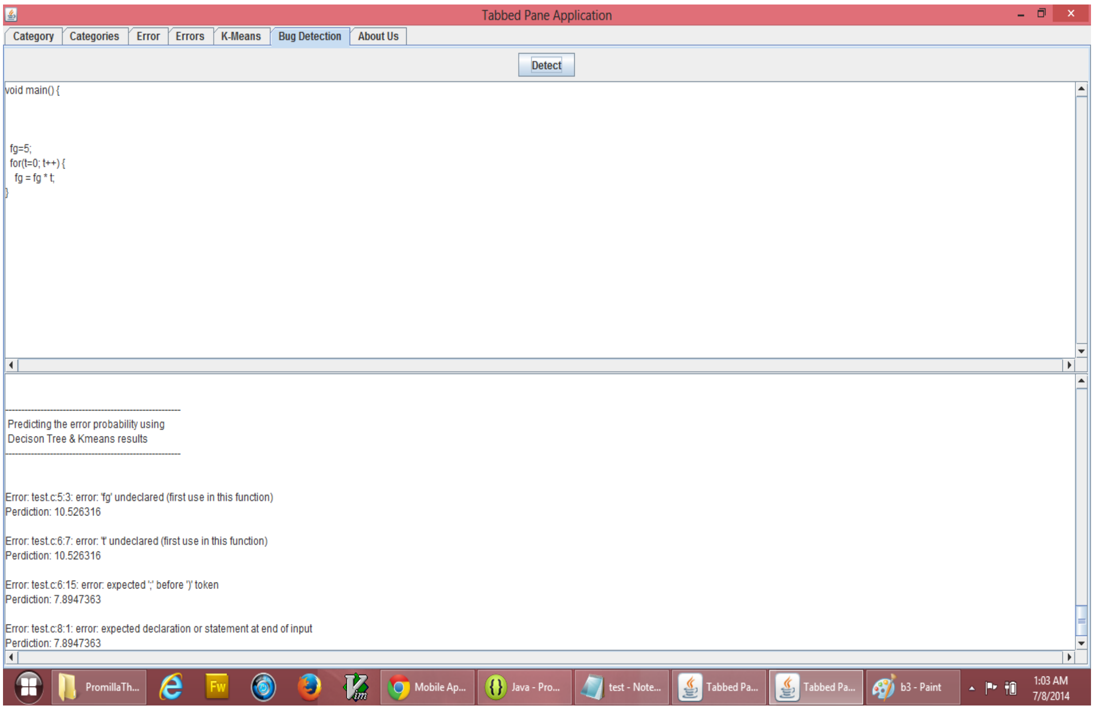
Result



```
void main() {
 int i=0;

int r;

 printf("%d", r);
 getch();

}
```

```
--------------------------------------------------------
Building C source code file...
--------------------------------------------------------

1) Analyzing error
  collect2: error: ld returned 1 exit status
  (*) Error classified successfully
  (*) Error category: 48
  (*) Error category description: [linker error]

--------------------------------------------------------
```

Fig 1.10



```
void main() {



 fg=5;
 for(t=0; t++) {
  fg = fg * t;
}
```

```
--------------------------------------------------------
 Predicting the error probability using
 Decison Tree & Kmeans results
--------------------------------------------------------


Error: test.c:5:3: error: 'fg' undeclared (first use in this function)
Perdiction: 10.526316

Error: test.c:6:7: error: 't' undeclared (first use in this function)
Perdiction: 10.526316

Error: test.c:6:15: error: expected ';' before ')' token
Perdiction: 7.8947363

Error: test.c:8:1: error: expected declaration or statement at end of input
Perdiction: 7.8947363
```

fig 1.11

Conclusion

Clustering is one of the most essential steps in data mining. It is the process of grouping data items based on similarity between elements in a cluster and dissimilarities between clusters. In this paper we have provided an overview of the broad classification of clustering algorithms such as partitioning, hierarchical, density based and grid based methods.

According to my project the bug has been detect by using c code compiler called in java ,it show multiple

errors in program when we execute program by using bug detect tab after writing the code then by detect button we will get result.It will display number of errors,their typer of error like as syntax error,undefined variables etc. and they also show there category to match the type of category.

In this bug detection code it will display cluster by using K-Means cluster algorithm and also detect Prediction using decision tree with K-Means.

Future scope:: For future research work Better clustering algorithm can be used. And More languages can be analyzed like dot net,C++,Python etc.

REFERENCES

[1] V. Neelima, Annapurna. N, V. Alekhya, Dr. B. M. Vidyavathi "Bug Detection through Text Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May2013.radio centric overlay-underlay waveform," in Proc. 3rd

[2] Anuja Priyama*, Abhijeeta, Rahul Guptaa ,Anju Ratheeb, and Saurabh Srivastavab, " Comparative Analysis of Decision Tree Classification Algorithms "International Journal of Current Engineering and Technology , Vol.3, No.2 (June 2013)

[3] Masud Karim, Rashedur M. Rahman ," Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing", Journal of Software Engineering and Applications, 2013, 6, 196-206.

[4] Dharminder Kumar, Suman," Performance Analysis of Various Data Mining Algorithms: A Review", International Journal of Computer Applications (0975 – 8887) Volume 32– No.6, October 2011

[5] M.E. Çelebi, H.A. Kingravi, P.A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm", Expert Systems with Applications 40, 2013, pp. 200-210.

[6] S.Z. Selim, M.A. Ismail, "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality". IEEE Transactions on Pattern Analysis and Machine Intelligence 6(1), 1984,pp. 81-87

[7] Chen, M. S., Han, J., & Yu, P. S., "Data Mining: An Overview from Database Perspective", IEEE Trans. on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 866-883, 1996

[8] C. Rygielski, J.C. Wang, D.C. Yen, "Data mining techniques for customer relationship management", Technology in Society 24, 2002,pp. 483-502.

[9] B Data Mining: Concepts and Techniques, 3rd Edition, Jiawei Han and Micheline Kamber, Jian Pei, 2007.

[10] Survey of Clustering Data Mining Techniques, Pavel Berkhin, 2002.

[11] Some Methods for classification and Analysis of Multivariate Observations, J. B. MacQueen, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967

[12] Clustering by means of medoids, L Kaufman and P Rousseeuw, Statistical Data Analysis Based on the L1-Norm and Related Methods, 19