SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 3
ISSN 2054-7390

Data Mining and Other Data Base Techniques for Ph.D Thesis Preparation

[1]Srinatha Karur, [2]M.V.Ramana Murthy
[1]*Orabyte Software Solutions, RTC X Roads, Hyderabad, India;*
[2]*School of Computer Science & Mathematics, Osmania University, Hyderabad, India;*
karurdori@gmail.com; mv.rm50@gmail.com

ABSTRACT

The authors in this paper present the role of Data mining and Data base techniques for estimate the quality of thesis or dissertation at Research level. The Doctorate Research consists of various components which are highly defined by University Research Committee or any other concerned authorities. For all general cases the thesis book consists of different chapters with different aims. Each and every chapter has its own identity and constantly has relation with previous chapters. The authors use different Data mining and Data base techniques for determine the correlation between different entities which are involved in the thesis such as page numbers, references, diagrams, equations , graphs, different concepts covered in the thesis book.

Keywords: Data mining techniques, SQL, thesis preparation, relation between entities.

1 Introduction

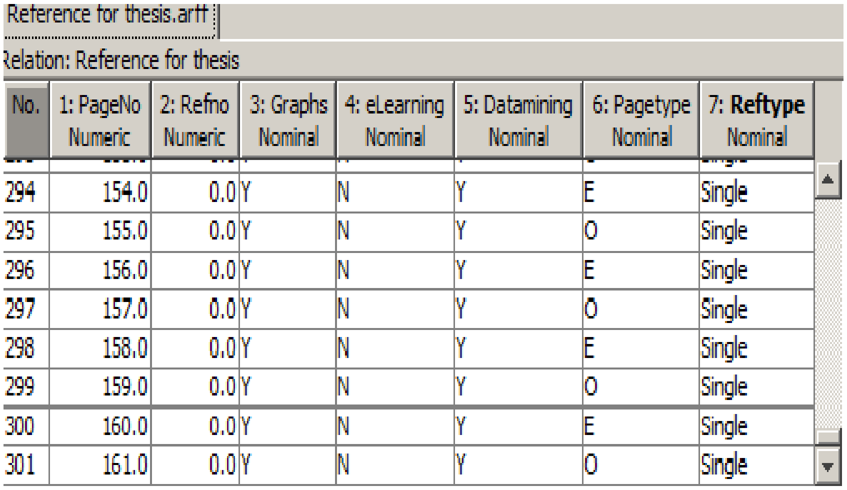
Education data mining become more dominant area and most consistent domain. Education Data mining deals about not only deals about the relation and scope between different Education systems and also how we can implement and do research as per needs of enhanced Education system. The enhanced systems generally deal about the higher versions of available systems. For example e-Learning is the enhanced version of Distance Education system. The Distance Education system is enhanced version of traditional or regular system. All enhanced versions are generally convergent as per static needs and once again fired or executes when requirement generated. It is like client server process when client requests server executes and vice versa. The PhD thesis are highest level of code execution or conduct for confer the degree by the University globally. We can use Data mining techniques either Supervised or Unsupervised or Hybrid methods or Semi supervised or Semi Unsupervised (Partial clusters). Semi unsupervised leads partial clusters and more inconsistent results are generated [1]. The author discussed about the semi unsupervised and semi supervised methods in [1] and defines the role and nature of label, unlabeled, and little label and little unlabeled. The authors published different papers on thesis preparation which covers all methods of data mining except Principle Component Analysis (PCA) since they need more knowledge on Mathematical and Statically concepts. Moreover PCA are used to convert orthogonal correlated to un correlated variables which is strictly out of scope of thesis [2]. General real time application or problem deals or gives important to estimate the correlation between available entities but not its negation values. Mainly the authors used Naïve Bayes method, Linear Regression Analysis, Rule based decision trees, Different probability distributions, nonlinear

DOI: 10.14738/tmlai.33.1227
Publication Date: 30th May, 2015
URL: http://dx.doi.org/10.14738/tmlai.33.1227

equations (Log and Exponential Curves), confusion matrix, Geni index, lift, outlier's estimation etc. as a part of supervised methods. The more information about data preparation and implementation details are available in [3, 4, 5, 6, 7, 8]. In these publications the authors are implemented the required phases very successfully and integrated the all phases of thesis for final submission. The authors are used most of the time Hierarchical clusters only on the basis of easy understanding. Since most of the Research community using Agloromative (Bottom to Top) Hierarchical clusters only. K-means also used frequently and EM also used as per context. The authors used very popular Data mining tools such as Tanagra, Weka, R, Orange, Rapid miner, as freeware tools and MS-SQL 2008 R2 as licensed software[8]. The author's main aim of this paper is to estimate the relation between page numbers and number of references in thesis book which is available as final copy for final submission to University as a vital part of course. Meanwhile the authors observed various parameters such as the role of supervised methods and unsupervised methods, results, analysis, tables etc. are available as a part of thesis and how they are related with each other..

2 Data preparation and Experiments

The authors use Weka for main Data mining processing and Tanagra for even Statistics events also. Microsoft Excel is used for find out the linear and other relationship between the defined or available variables. The authors use both continuous and discrete variables for Data preparation and implementation purpose. The below figure shows data is successfully loaded into Weka in .arff form with 7 fields and 301 instances are as follows.



Reference for thesis.arff

Relation: Reference for thesis

No.	1: PageNo Numeric	2: Refno Numeric	3: Graphs Nominal	4: eLearning Nominal	5: Datamining Nominal	6: Pagetype Nominal	7: Reftype Nominal
294	154.0	0.0	Y	N	Y	E	Single
295	155.0	0.0	Y	N	Y	O	Single
296	156.0	0.0	Y	N	Y	E	Single
297	157.0	0.0	Y	N	Y	O	Single
298	158.0	0.0	Y	N	Y	E	Single
299	159.0	0.0	Y	N	Y	O	Single
300	160.0	0.0	Y	N	Y	E	Single
301	161.0	0.0	Y	N	Y	O	Single

Figure 1: Shows data is successfully loaded

Page numbers and reference numbers are numeric whereas remaining are character type and we can prepare the data as per needs and method. The weka tool gives various distributions with respect to different field values are as follows.

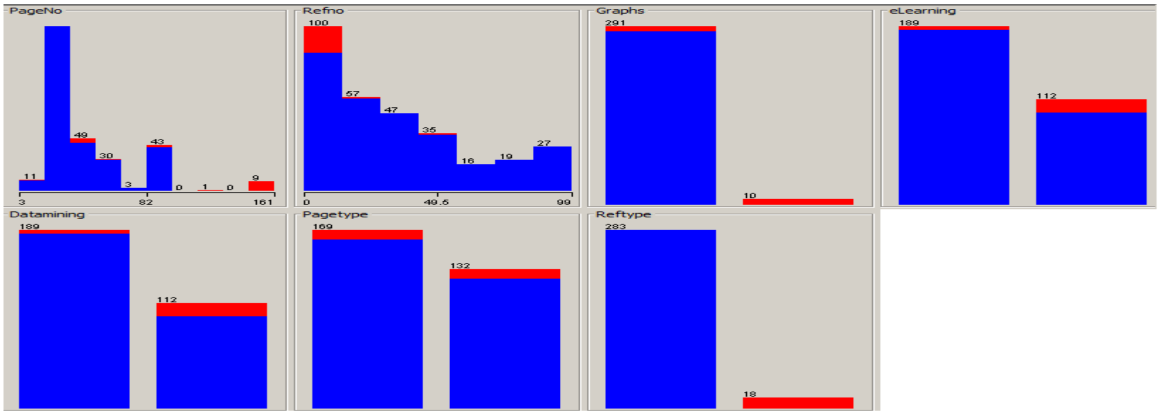


Figure 2: Shows success and failure rates for different classes in data

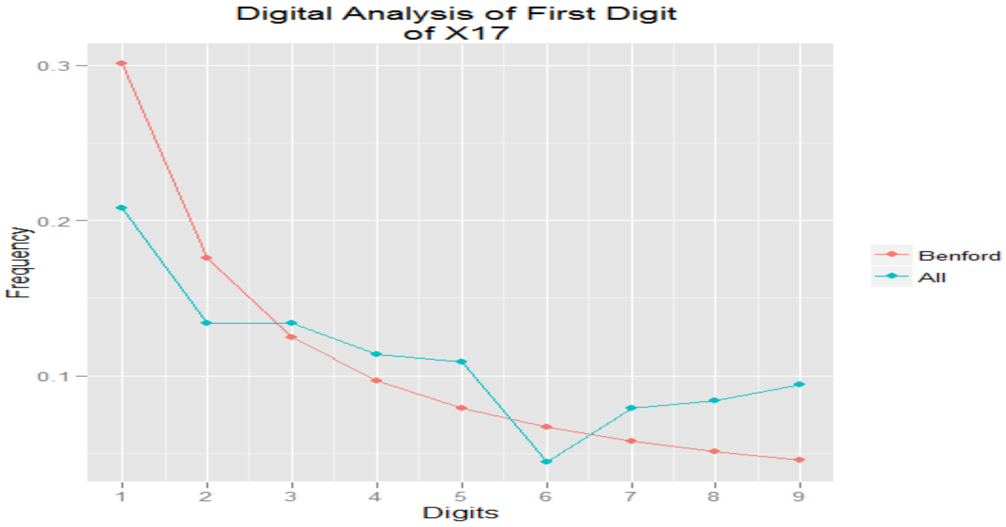


Figure 3 : Shows Benford graph for x and y components

Benford's law, also called the First-Digit Law, refers to the frequency distribution of digits in many (but not all) real-life sources of data. In this distribution, 1 occurs as the leading digit about 30% of the time, while larger digits occur in that position less frequently: 9 as the first digit less than 5% of the time. Benford's law also concerns the expected distribution for digits beyond the first, which approach a uniform distribution. The mathematical equation of Benford law is as follows.

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}\left(1+\frac{1}{d}\right). \qquad (1)$$

The quantity P(d) is proportional to the space between d and d + 1 on a logarithmic scale. n extension of Benford's law predicts the distribution of first digits in other bases besides decimal; in fact, any base b ≥ 2. For example the linear regression analysis only first two field values are enough and for Naïve Bayes only classes or attributes are enough. The authors used MS-Excel for estimate the linear and higher degree relations which are as follows and have been shown in below graph form.

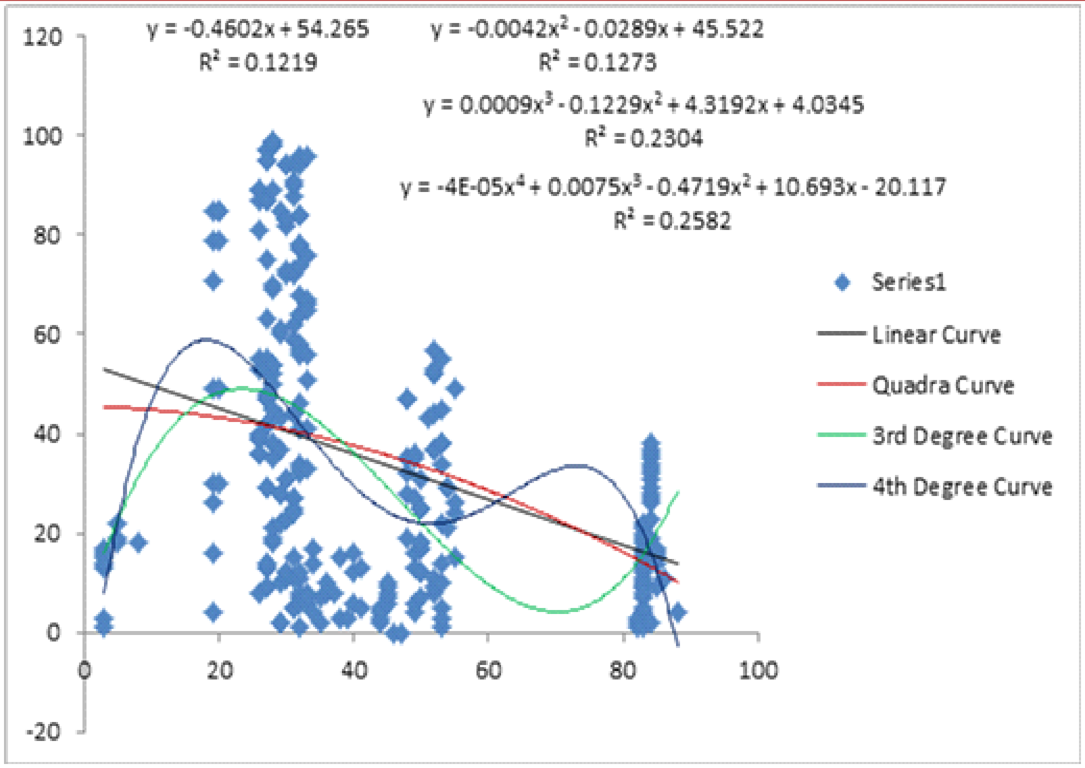


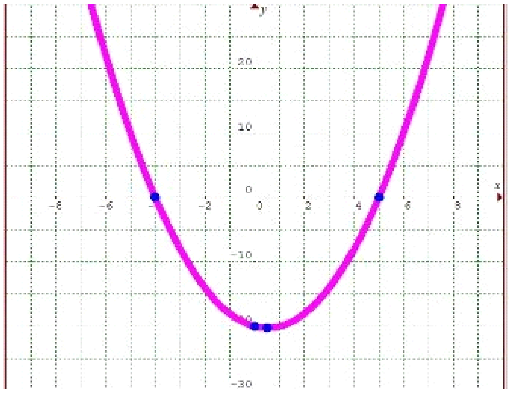
Figure 4: linear and higher degrees for page number and number of references

The authors used to find out the higher degree of relations between page number and number of references which are used during thesis writing. The authors observed that linear and secondary orders are formed almost straight lines and especially linear curve is highly intersects at x and y axis with in first quadrant as shown in the figure. The higher orders > 2 are strictly curve nature is as shown in the figure-3. For residuals and standard error estimation the authors are used curve expert software with numeric values of first and second fields of given data base with 301 instances are shown in the figure-1. The general form of non linear function is given by f(x) = anxn + an-1xn-1 + an-2xn-2 +…… + a1x1 + a0 where a0, a1……. an are stables. In this non linear functions, an is a primary co –efficient and anxn is the principal term. The greatest degree of non – linear function is greater than or similar to 2.A quadratic function is in the form y = ax2 + bx + c, where c ≠ 0 is a non-linear equation. Similarly, a cubic function y = ax3 + bx2 + cx + d, where a≠ 0 is a non-linear equation. Non-linear functions are those which do not form a straight line when graphed. One of the functions which are not a linear function and cannot be a complete linear function by transforming the Y variable.

There are three nonlinear functions normally used in mathematics as follows,

- Exponential function
- Quadratic function
- Logarithmic function

The meaning of the non-linear functions cannot be overstated since without them, thus without graphing it would not be a function. The original testing in the field of simulated equations failed since there was no clear understanding of the importance of non linearity in the output point.



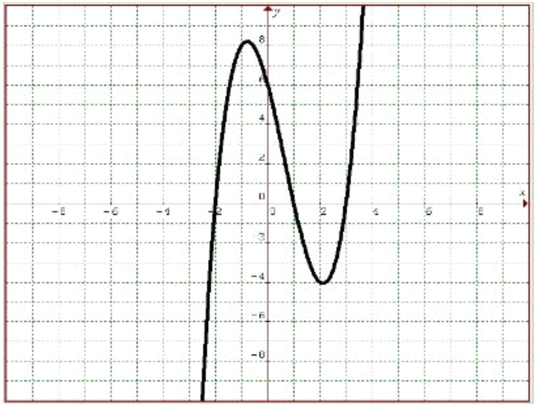
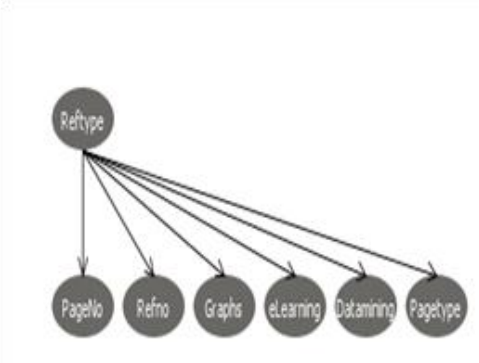


Figure 5: Non linear form for degree 2

Figure6: Non linear form of degree>2

The authors observed the same data for NavieBayes Networks and NavieBayes the experiment is repeated for different values and dimensions. All values are recorded and mentioned in Results section. The authors observed the tree in both Naive layout and priority layout. The figures are as follows.



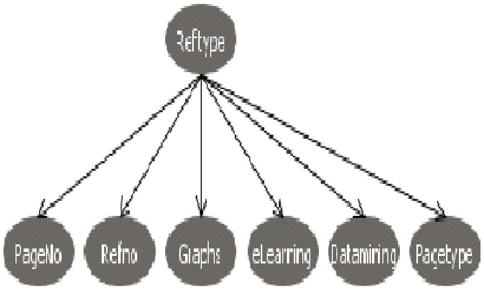


Figure 7: Shows Navies layout and each node has prob>0.2

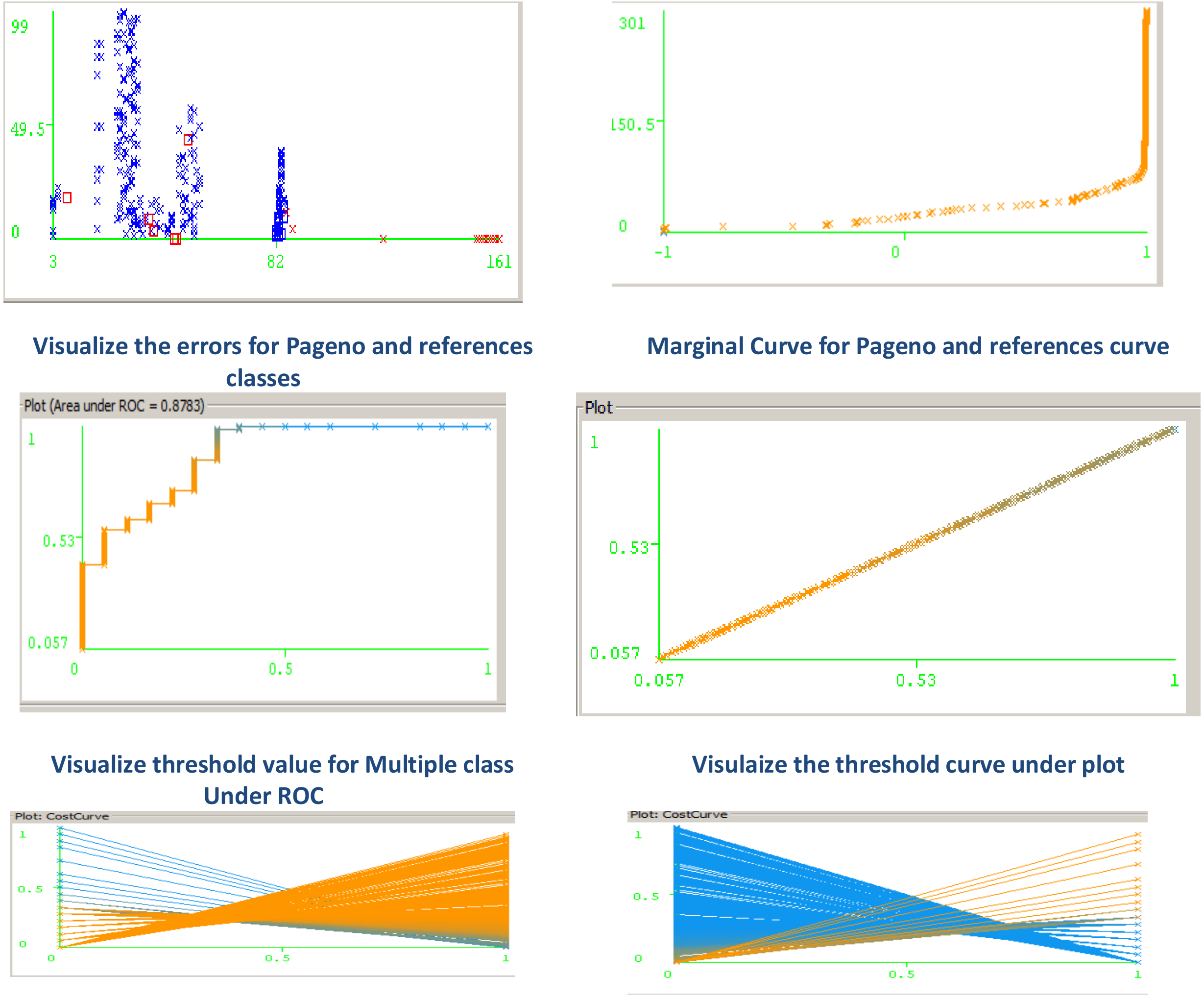
Figure 8: Shows priority model and minimum probability is 0.3

All confusion matrices are available in Results and analysis section and the authors are observed that there is little variation in confusion matrices of Navie and priority models. Due to the format of the paper the authors did not present the confusion matrix values in this section and it is available along with other experiment results. The authors test the data with Naive bayes model consists of the following things.

- Visualization Margin curve
- Visualize threshold curve
- Cost/Benefit analysis
- Cost curve analysis.

The authors observed the following things during the data testing and analysis for Naive Bayes. The authors test the data with this constraint

- Along the x-axis False positive rate
- Along the y-axis True positive rate



Visualize the errors for Pageno and references classes

Marginal Curve for Pageno and references curve

Visualize threshold value for Multiple class Under ROC

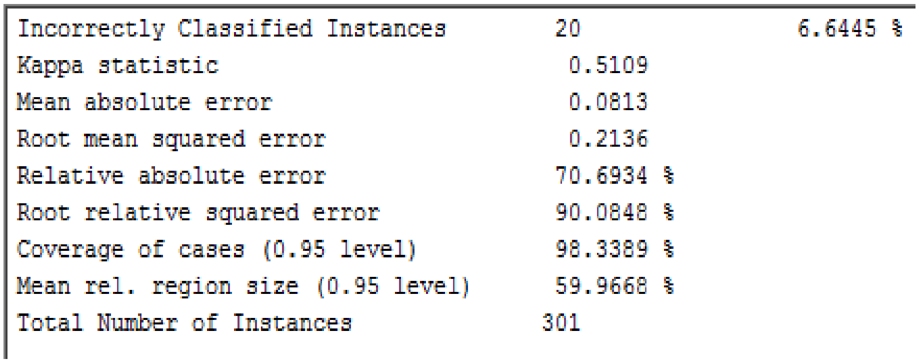
Visulaize the threshold curve under plot

For multiple class cost curve

For single class Cost curve

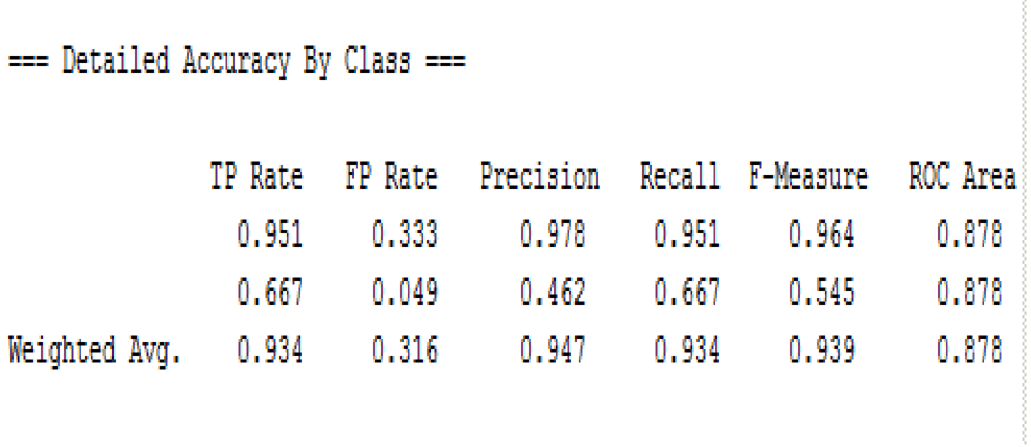
Figure 9: Shows Naive Bayes modeling with factors

The authors tested the data with various properties are shown in the figure as follows. The mathematical statistics for the given data is as follows.



```
Incorrectly Classified Instances        20               6.6445 %
Kappa statistic                          0.5109
Mean absolute error                      0.0813
Root mean squared error                  0.2136
Relative absolute error                 70.6934 %
Root relative squared error             90.0848 %
Coverage of cases (0.95 level)          98.3389 %
Mean rel. region size (0.95 level)      59.9668 %
Total Number of Instances              301
```

Figure 10: Statistics for Page no and reference



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.951     0.333     0.978     0.951     0.964      0.878
                 0.667     0.049     0.462     0.667     0.545      0.878
Weighted Avg.    0.934     0.316     0.947     0.934     0.939      0.878
```

Figure 11: Different parameters for Naïve Bayes.

Table 1: Shows nature of single and multiple, classes graph nature for given data

S. No	Property	Single class	Multiple class
1	Precision	Depend	depend
2	Recall	Strictly diagonal	Diagonal
3	Fall out	Depend	depend
4	F-Measure	Curve on y-axis	Almost diagonal
5	Sample size	depend	diagonal
6	Lift	Zigzag line on y-axis	depend
7	Curve nature	depend	Generally diagonal

The authors observed the relation between various intervals for Single and multiple classes are as follows. Empty cells indicate any type of nature even straight line passing through origin. For linear and non linear relations along with R^2 values are as follows (all r2 values are >0)

$$y=1.103x+0.717 \tag{2}$$

$$R^2 = 0.621 \tag{3}$$

$$y = -2.150x^2 + 3.324x + 0.628 \tag{4}$$

$$R^2 = 0.644 \tag{5}$$

$$y = 121.3x^3 - 135.4x^2 + 15.39x + 0.542 \tag{6}$$

$$R^2 = 0.730 \tag{7}$$

The authors observed the following points are as their observation

- All r2 values are >0
- All x-coefficients are >0
- All constants are >0

- For quadratic equation factors are x1>0 and x2=1.7162

The authors used online tool for solve the quadratic equation [9].

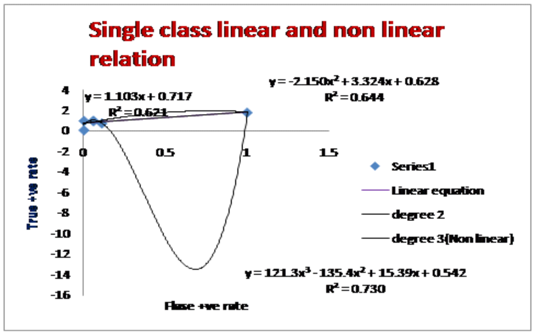


Figure 12: Shows linear and non linear for Single class for Naive Bayes

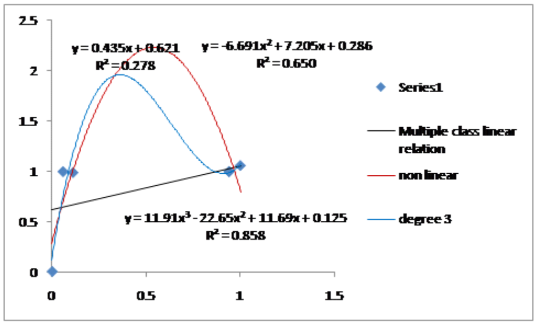
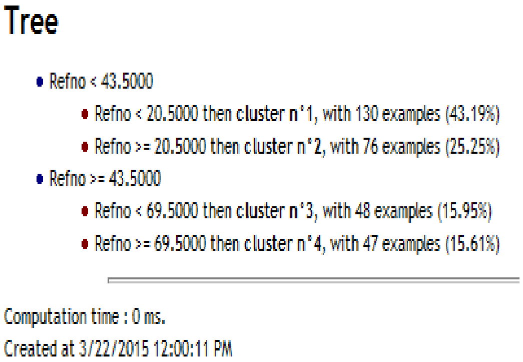


Figure 13: Shows linear and non linear for Multiple classes Naïve Bayes



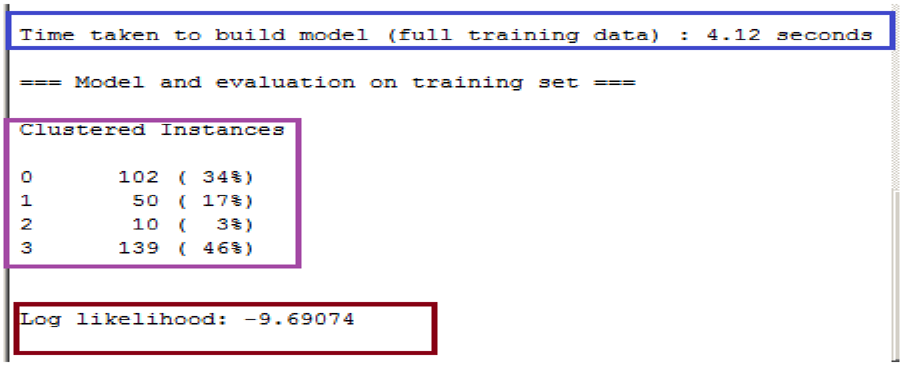
Tree

- Refno < 43.5000
  - Refno < 20.5000 then cluster n°1, with 130 examples (43.19%)
  - Refno >= 20.5000 then cluster n°2, with 76 examples (25.25%)
- Refno >= 43.5000
  - Refno < 69.5000 then cluster n°3, with 48 examples (15.95%)
  - Refno >= 69.5000 then cluster n°4, with 47 examples (15.61%)

Computation time : 0 ms.
Created at 3/22/2015 12:00:11 PM

Figure 13A: Tree rules for given data

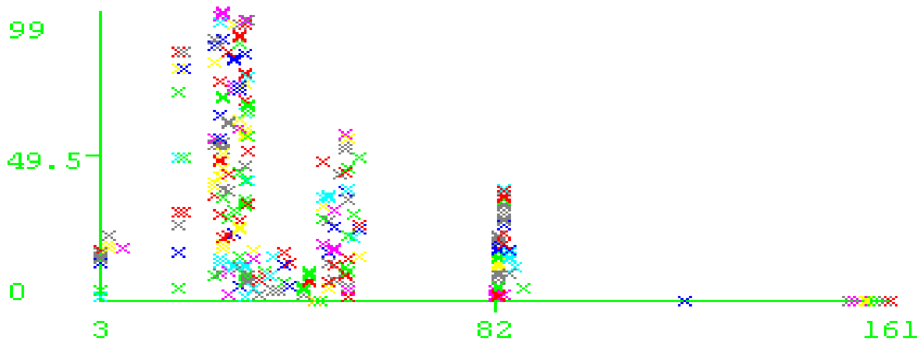
The authors tested the data for Unsupervised and semi unsupervised methods such as EM method and the diagrams are as follows.



```
Time taken to build model (full training data) : 4.12 seconds

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

Clustered Instances

0       102 ( 34%)
1        50 ( 17%)
2        10 (  3%)
3       139 ( 46%)


Log likelihood: -9.69074
```

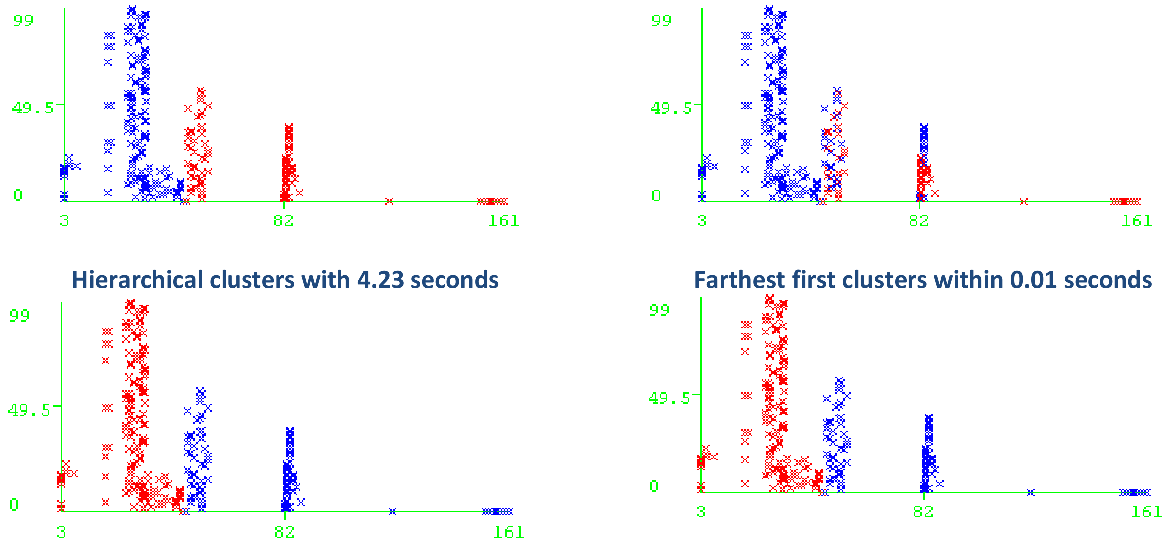
Figure 14: EM implementation with Weka with log likely hood -9.7

From the above figure it is observed that four clusters are formed and second cluster consists of only 3% of instances which forms very small cluster and cluster 3 is a big cluster and has 46% of instances. Both are at extreme values . For many applications, the natural logarithm of the likelihood function, called the log-likelihood, is more convenient to work with. Because the logarithm is a monotonically increasing function, the logarithm of a function achieves its maximum value at the same points as the function itself, and hence the log-likelihood can be used in place of the likelihood in maximum likelihood estimation and related techniques. Finding the maximum of a function often involves taking the derivative of a function and solving for the parameter being maximized, and this is often easier when the function being maximized is a log-likelihood rather than the original likelihood function. It is observed that from the above figure minimum and maximum occurrences are one by one and intermediate values are formed randomly. The four clusters formation is as shown in the figure



Cobweb method shows 4 clusters are formed with 0.016 seconds

It is observed that from the figure between [3,82] interval maximum events are occurred and after this interval almost negligible events are occurred and these events or objects are called idle objects and independent on clusters and we can say deviate from clusters almost. The authors repeat the experiment for all remaining unsupervised methods and the noted the contents is as follows.



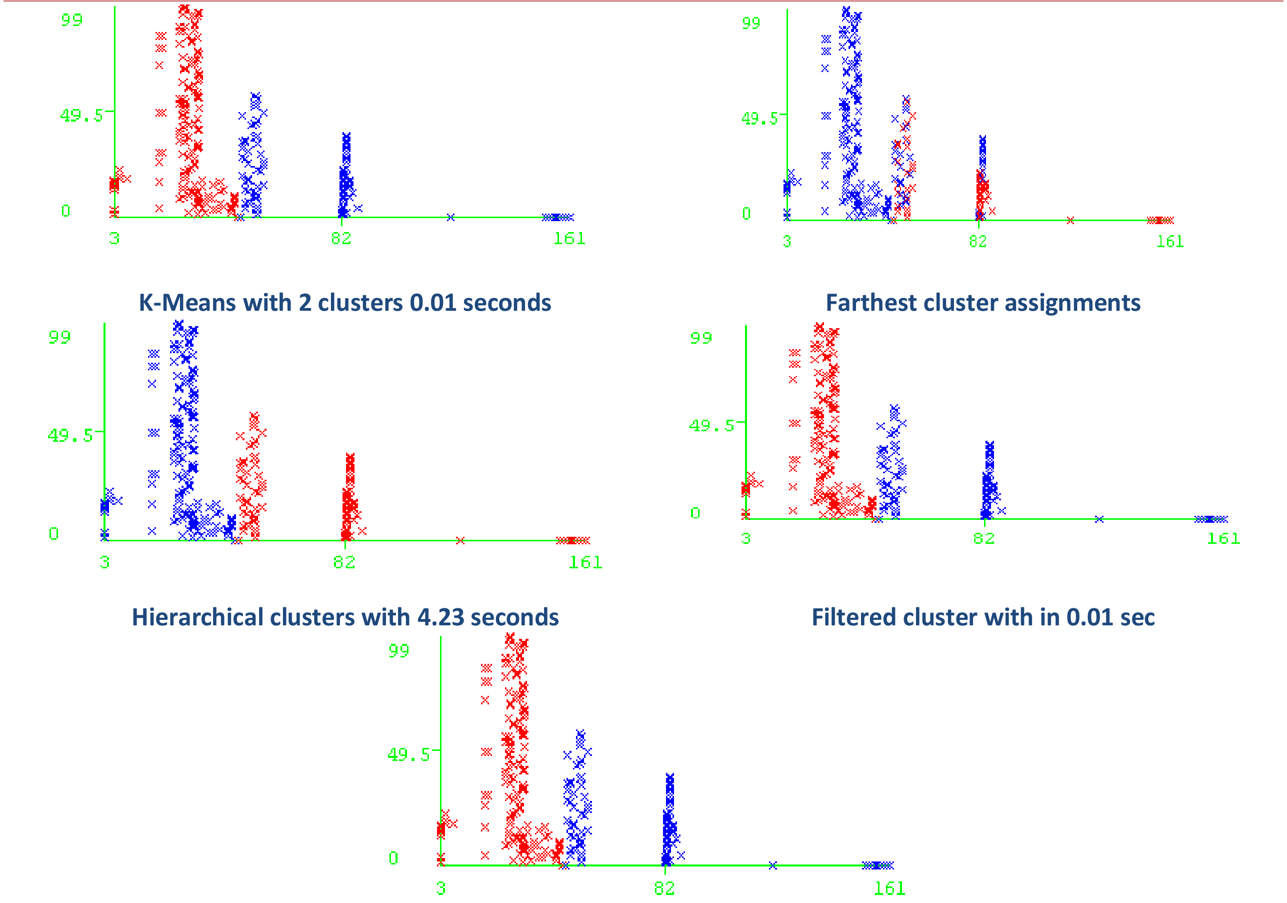
Hierarchical clusters with 4.23 seconds

Farthest first clusters within 0.01 seconds

Filtered cluster with in 0.01 sec

Density based clusters with -10.25311 likely hood functions with 0.04 seconds

Figure 15: shows different methods of Clusters



K-Means with 2 clusters 0.01 seconds

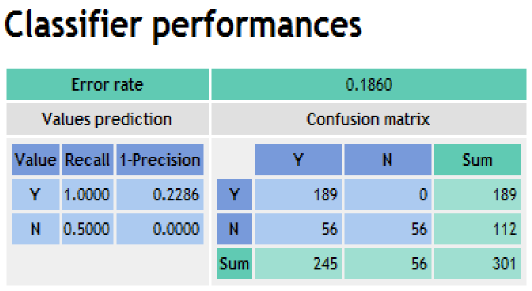
Farthest cluster assignments

Hierarchical clusters with 4.23 seconds

Filtered cluster with in 0.01 sec

Density based clusters with -10.25311 likely hood function with 0.04 seconds

Figure 16: Shows different clusters assignments



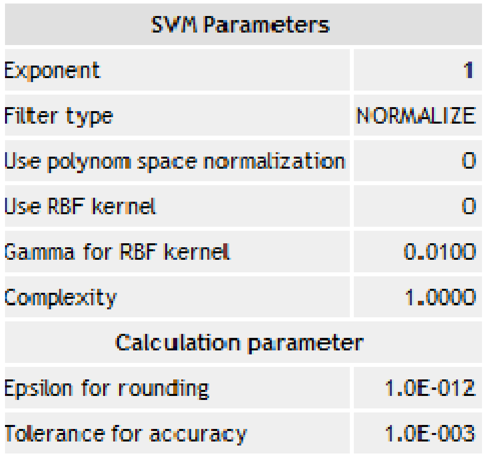
Classifier performances

Error rate	0.1860
Values prediction	Confusion matrix

Value	Recall	1-Precision
Y	1.0000	0.2286
N	0.5000	0.0000

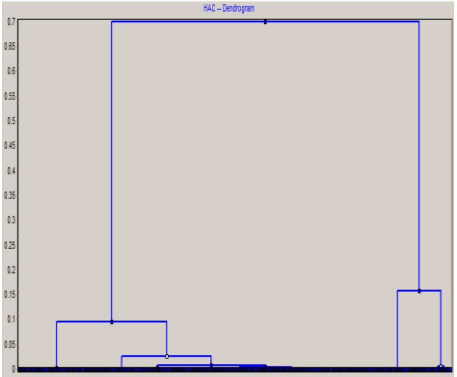
	Y	N	Sum
Y	189	0	189
N	56	56	112
Sum	245	56	301

Figure 17B: Shows cluster assignments and SVM parameters

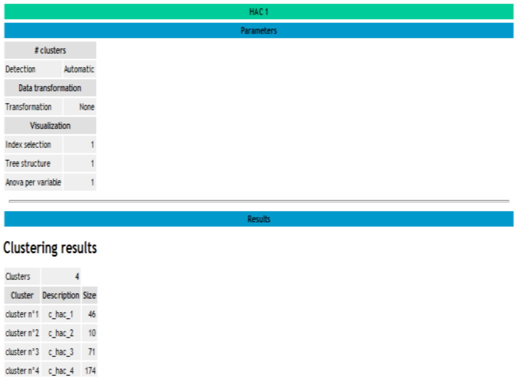


SVM Parameters	
Exponent	1
Filter type	NORMALIZE
Use polynom space normalization	0
Use RBF kernel	0
Gamma for RBF kernel	0.0100
Complexity	1.0000
Calculation parameter	
Epsilon for rounding	1.0E-012
Tolerance for accuracy	1.0E-003

Figure 17A: Shows SVM parameters



Hierachical clusters with TANAGRA with height=0.7



HAC 1

Parameters

# clusters	
Detection	Automatic

Data transformation	
Transformation	None

Visualization	
Index selection	1
Tree structure	1
Anova per variable	1

Results

Clustering results

Clusters	4	
Cluster	Description	Size
cluster n°1	c_hac_1	46
cluster n°2	c_hac_2	10
cluster n°3	c_hac_3	71
cluster n°4	c_hac_4	174

Formed 4 clusters with maximum and minimum instances

Figure 18: Shows HR with 4 clusters

3 Results and Analysis

The authors tested and observed the data between no of pages and number of references is as shown in figure-1. The authors repeat the HC repeatedly for different types of distances and links then the results are as follows. The authors used R software with Rattle GUI for this purpose and the tool snapshot is not available in this paper.

Table-2 shows HC for different Distance methods and links

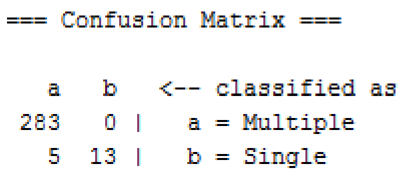
S. No	Distances	Wards	Complete	Single	Average	Mequitty	Median	Centroid
1	Euclidian	2000	90	20	60	60	40	40
2	Maximum	2300	42	25	50	60	40	42
3	Manhattan	4000	150	30	80	80	65	50
4	Canberra	40	2.0	1.0	1.5	1.75	2	1.5
5	Binary	0	0	0	0	0	0	0
6	Pearson	30	0.8	0.02	0.3	0.3	0.3	0.3
7	Correlation	190	2.0	0	1.75	1.75	2	1.8
8	Spearman	200	2.0	2.0	2.0	2.0	2.0	2.0

The authors used supervised methods also for estimate the classifiers of a given problem. More details are available in [11]. Different data mining tools are available for applying these evaluation methods. For all popular tools such as Weka, Tanagra, Oranage, Rapid miner, R with Rattle are used CONFUSION MATRIX as common evaluation methods. For more details of this implementation by these tools are available in their respective documentation. The tool R consists of not only command prompt and also lot of GUI tools

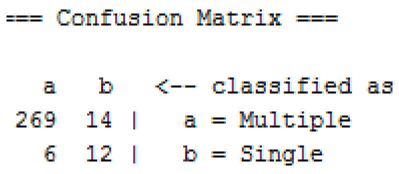
for implementation as highly user friendly. For more details of tools of R is available in http://www.linuxlinks.com/article/20110306113701179/GUIsforR.html. The authors applied different evaluation methods for Naïve Bayes are as follows.

- Visualization Margin curve
- Visualize threshold curve
- Cost/Benefit analysis
- Cost curve analysis.

All the nature of graphs are tabulated in table-1. Confusion matrix for Naïve net and Naïve Bayes , and Naïve Bayes Update is as follows.

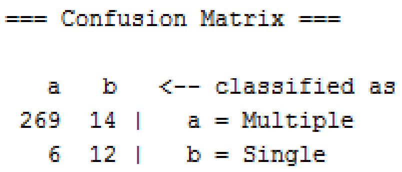


```
=== Confusion Matrix ===

   a   b   <-- classified as
 283   0 |   a = Multiple
   5  13 |   b = Single
```



```
=== Confusion Matrix ===

   a   b   <-- classified as
 269  14 |   a = Multiple
   6  12 |   b = Single
```

Naive Bayes update table



```
=== Confusion Matrix ===

   a   b   <-- classified as
 269  14 |   a = Multiple
   6  12 |   b = Single
```

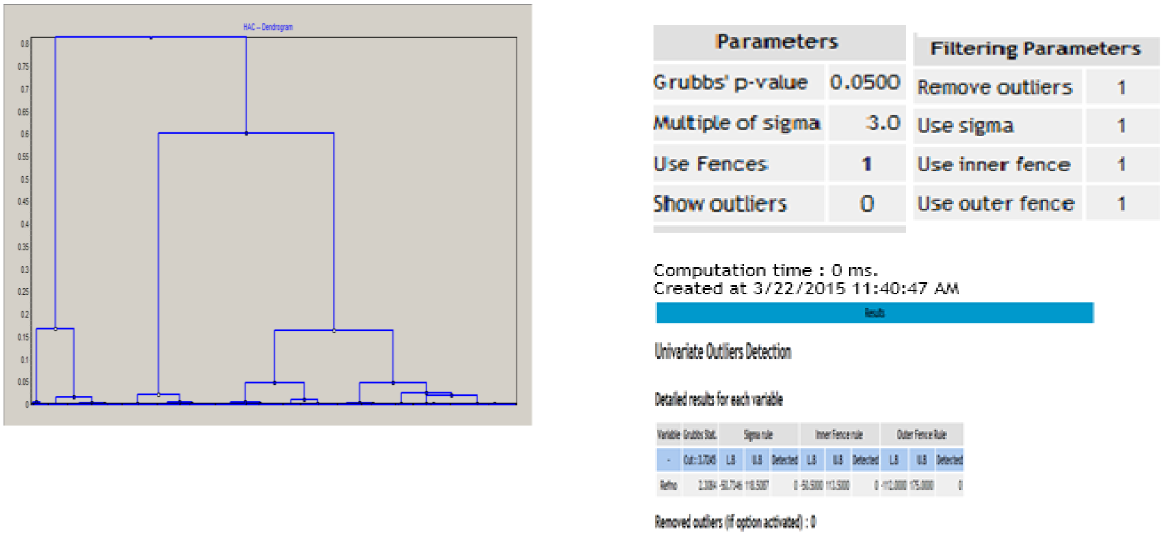
Naive Bayes confusion matrix

Figure 19: Shows confusion matrix for Naïve net, Naïve Bayes and Update Naïve models.

The authors note the generated output of Naïve Bayes Evaluation methods are as follows. The evaluation methods are mentioned in Table-3

S. No	Name	Single Class IV	Multi Class IV
1	Precision	[0.06,1]	[0.94,1]
2	Recall	[0.056,1]	[0.057,1]
3	Fall out	[0,0.94]	[0,0.06]
4	FMeasure	[0.11,0.76]	[0.11,0.99]
5	Sample size	[0.0033,1]	[0.94,1]
6	Lift	[1,11.2]	[1,1.06]

The cost benefit analysis for Naïve Bayes method is -11.2957 and 97.6744 with respect to Max cost and min cost where along the x axis Sample size and along the y axis cost/benefit is available. More details and mathematical model of Supervised Vector Machine are available in [11]. SVM mathematical modeling is like Linear Programming problem model and its details are out of scope. The authors finally tested for clustering and tree rules. The output is as follows. Finally the authors used to find out the outliers estimation for given or prepared data. All the above aims are available as follows. For this the authors used TANAGRA software. It is observed that 0 outliers are found for Univariate from below figure-21.



Parameters	
Grubbs' p-value	0.0500
Multiple of sigma	3.0
Use Fences	1
Show outliers	0

Filtering Parameters	
Remove outliers	1
Use sigma	1
Use inner fence	1
Use outer fence	1

Computation time : 0 ms.
Created at 3/22/2015 11:40:47 AM

Results

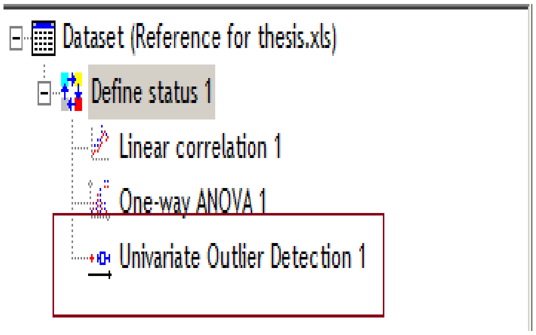
Univariate Outliers Detection

Detailed results for each variable

Variable	Grubbs Stat.	Sigma rule			Inner Fence rule			Outer Fence Rule		
-	Crit : 3.7045	LB	UB	Detected	LB	UB	Detected	LB	UB	Detected
Refno	2.3094	-50.7346	118.5387	0	-50.5000	113.5000	0	-112.0000	175.0000	0

Removed outliers (if option activated) : 0

Figure 20: For both tree rules and HC



Dataset (Reference for thesis.xls)
Define status 1
Linear correlation 1
One-way ANOVA 1
Univariate Outlier Detection 1

Figure 21: Shows outliers estimation for given thesis

4 Conclusions

Using Data mining techniques it is possible to estimate the relation between to estimate the relation between various factors such as relation between number of pages and references used, how the chapters are distributed and how the topics are distributed, Statistical view of entire nature of distribution of information, how the diagrams are correlated and how many diagrams and tables are arranged in odd and even pages and their relation etc. we can find easily. We can also find the role of mathematical equations and its distribution throughout the thesis book. The authors are used only interpretation concept and did not test for orthogonal trajectories which mainly deals the Principal component analysis. The authors also estimated the outliers for given thesis and found that almost zero errors are available. The authors used only TANAGRTA software for outlier's estimation. We can examine these outliers nature and estimation using different free Data mining tools such as Weka, Orange, Rapid miner and R.

REFERENCES

[1] www.umiacs.umd.edu/~hal/docs/daume09sslnlp.pdf

[2] http://en.wikipedia.org/wiki/Principal_component_analysis.

[3] www.airccse.org/journal/ijaia/papers/4513ijaia02.pdf

[4] www.airccse.org/journal/ijaia/papers/4413ijaia12.pdf

[5] www.gssrr.org/index.php?journal=JournalOfBasicAndApplied..,

[6] www.ijecs.in/issue/v2-i10/16%20ijecs.pdf

[7] http://www.ijettcs.org/Volume2Issue6/IJETTCS-2013-12-10-061.pdf

[8] www.ijaiem.org/volume3issue5/IJAIEM-2014-05-29-093.pdf

[9] http://www.math.com/students/calculators/source/quadratic.htm

[10] http://www.bth.se/fou/forskinfo.nsf/0/c655a0b1f9f88d16c125714c00355e5d/$file/Lavesson_lic.pdf

[11] http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf