Comparative analysis of TOFEL IBT Result rate Among Students using K-Means Clustering

J.Emmanual Robin, G.Prabu

Abstract*—. *Data mining technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. This paper reveals the comparative analysis of the students with UG,PG,Other Students community. Before getting into the picture we have to know the basic concept of clustering technique. What is clustering analysis? Clustering analysis divides data into the groups (clusters) that are meaningful or useful or both. If meaningful groups are the goal, then the clusters should capture natural structure of data. This paper focuses to discover the comparative analysis of reading, writing, speaking, listening skills over the student's dataset such as

(a)PercentileMarkUG(b)PercentileMarkPG (c)PercentileMarkOther

Index Terms*— *K-Means Clustering, IBT(Internet Based Test),TOEFL(Test of English as a Foreign Language)

I. INTRODUCTION

In this paper, deals new series for analysis of competitive exams over worldwide. In the global economy we are all supposed to master the mysterious of English. But how do you know if your understanding of the idioms and prepositions that keep the language schools in business. In the world of business what is the quantified assessment of your level for the American universities? The answer is TOEFL (Test of English as a Foreign Language).

This paper going to discuss the relationship among the students to compare the result analysis of skills like reading, listening, writing, speaking. It mainly focus in the form of deep discovery of TOEFL test is success rate or failure occurs in the students community. This analysis will really help the arial route with future projections in the different categorical perspective. This Process is done in Weka tool.

II. DATA SETS USED

The datasets available in www.ets.org/toefl/ 2009

Table-1(Score Scale)

Serial. No	Section	Min	Max
1	Reading	0	30
2	Writing	0	30
3	Listening	0	30
4	Speaking	0	30
Total			120

Table-2(Percentile Score)

Serial No	Scale Score	Reading	Listening	Writing	Speaking	Total Score	Percentile Score
1	30	96	97	99	99	120	100
2	29	89	90	98	96	116	99
3	28	83	85	96	92	112	96
4	27	77	79	92	86	108	92
5	26	72	74	88	**	104	87
6	25	66	69	**	77	100	81
7	24	61	64	81	68	96	74
8	23	56	59	71	**	92	67
9	22	51	54	60	56	88	59
10	21	46	50	**	45	84	52

Note: ** Non-Existent Score for Writing and Speaking

Table-3(Percentile Score for UG Students)

Serial No	Scale Score	Reading	Listening	Writing	Speaking	Total Score	Percentile Score
1	30	97	97	99	99	120	100
2	29	91	90	97	96	116	99
3	28	86	84	94	92	112	96
4	27	82	79	90	86	108	92
5	26	77	74	84	**	104	86
6	25	73	69	**	78	100	80
7	24	69	64	77	69	78	100
8	23	64	59	67	**	92	68
9	22	60	54	56	58	88	62
10	21	55	50	**	48	84	55

Note: ** Non-Existent Score for Writing and Speaking

Manuscript received January 10, 2012.

J.Emmanual Robin, Asst Professor, Department of Computer Applications,Jayaram College of Engg &Tech, Tiruchirappalli Tamil Nadu, India e-mail: erobinjoseph@gmail.com).

G.Prabu, Associate Professor, Department of Science and Engineering, Jayaram College of Engg &Tech, Tiruchirappalli Tamil Nadu,India e-mail: vgprabhu.samy@gmail.com)

Table-4(Percentile Score for PG Students)

Serial No	Scale Score	Reading	Listening	Writing	Speaking	Total Score	Percentile Score
1	30	97	97	99	99	120	100
2	29	91	90	97	96	116	99
3	28	86	84	94	92	112	96
4	27	82	79	90	86	108	92
5	26	77	74	84	**	104	86
6	25	73	69	**	78	100	80
7	24	69	64	77	69	78	100
8	23	64	59	67	**	92	68
9	22	60	54	56	58	88	62
10	21	55	50	**	48	84	55

Note: ** Non-Existent Score for Writing and Speaking

Table-5(Percentile Score for Other Students)

Serial No	Scale Score	Reading	Listening	Writing	Speaking	Total Score	Percentile Score
1	30	96	96	100	99	120	100
2	29	88	89	98	97	116	99
3	28	82	82	96	93	112	96
4	26	72	72	88	**	104	85
5	26	72	72	88	**	104	85
6	25	67	67	**	77	100	79
7	24	63	62	80	67	96	72
8	23	58	58	69	**	92	65
9	22	54	53	59	56	88	59
10	21	50	49	**	46	84	53

Note: ** Non-Existent Score for Writing and Speaking

III. K-MEANS ALGORITHM

The K-Means algorithm takes the input parameter k and partitions a set of n objects into k clusters so that the resulting intra cluster is low. Cluster similarly measured in regard to the mean value of the objects in cluster, which can be viewed as the cluster's centroid.

First it randomly selects k of the objects. Each of which initially represents a cluster mean or center. For each of remaining objects ,an object assigned to the clusters to which it is most similar, based on the distance between the object and cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function coverages.Typically SSE(Sum of Squared Error) defined as:

$$SSE=\sum_{i=1}^{K} \sum_{x \in C_i} dist(C_i,x)^2 \text{ where } C_i=1/m_i \sum_{x \in C_i} x$$

Where C_i- Cluster Index-no of Points-Summation of Attribute Set, m_i-mean of Cluster

Example: Clusters is having three Two Dimensional Points (1,1),(2,3) and (6,2) is(1+2+6/3,1+3+2/3)=(3,2)

Step 1: Select K Points as initial centroids
Step 2: **repeat**
Step 3: Form K Clusters by assigning each point to its closest centroid.
Step 4: Recompute the centroid of each cluster
Step 5: until Centroids do not change

IV. METHODOLOGY

A. Process method

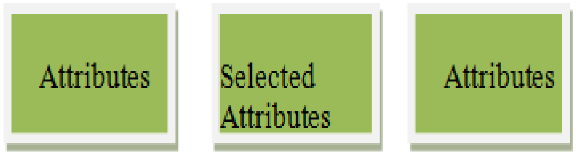

Fig-1- Preprocessed Data or Cleaning the Data

In this process raw data is cleaning and converted into the .csv format.

B. Preprocess Steps

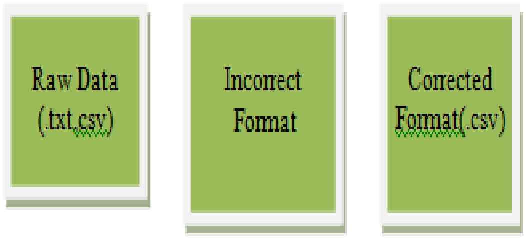

Fig-2 Weka Explorer tool Pre Process Steps

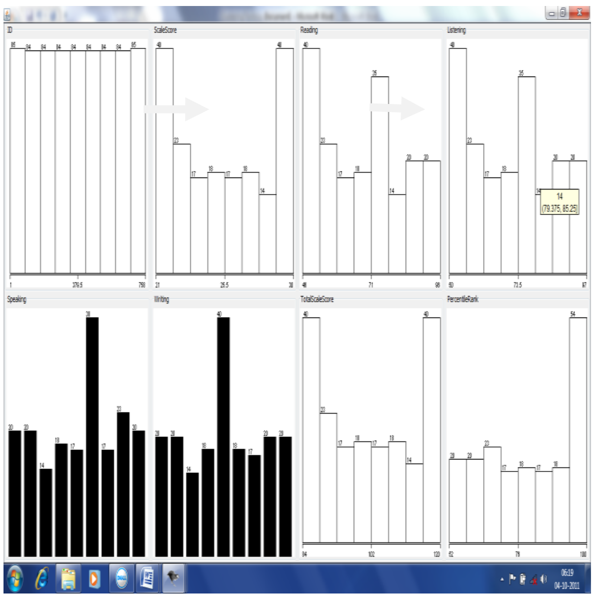

Fig-3 Weka Explorer tool Visualize Attributes

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. MODEL TRAINING AND TEST SET

Number of iterations: 3
Within cluster sum of squared errors:
393.29773823005354
Missing values globally replaced with mean/mode
Cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(758)	(40)	(718)
ID	379.5	86.4	395.8287
Scale Score	25.385	21.5	25.6015
Reading	69.1176	48.5	70.2663
Listening	71.5348	52	72.623
Speaking	**	**	**
Writing	**	56.0	**
TotalScaleScore	101.5401	86	102.4059
Percentile Rank	79.8717	55.5	81.2294

TEST EVULUATION
Number of iterations: 3
Within cluster sum of squared errors:
209.65116528306973
Missing values globally replaced with mean/mode
Cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(454)	(38)	(416)
ID	386.2269	91.1842	413.1779
Scale Score	25.0541	28.3947	24.7489
Reading	67.2432	85.5526	65.5708
Listening	69.7658	87.0526	68.1867
Speaking	**	98.0	**
Writing	**	96.0	**
TotalScaleScore	100.2162	113.5789	98.9956
Percentile Rank	78.2252	96.5263	76.5535

Clustered Instances

Fig 4-Percentile Score Instances and Centeroids

In this model training set gives the picture of cluster instances and centroids of each attributes.

VI. RESULTS AND DISCUSSIONS

Fig 4-Percentile Score Graph for UG Students (Visualize Attributes)

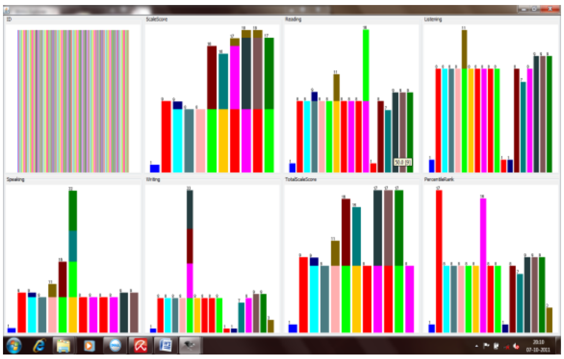

Fig 5-Percentile Score Graph for UG Students(Visualize Attributes)

Number of iterations: 2
Within cluster sum of squared errors:
863.0000000000001
Missing values globally replaced with mean/mode
Cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(136)	(104)	(32)
ID	ID	ID	6.0
ScaleScore	23.0	23.0	21.0
Reading	55.0	86.0	40.0
Listening	74.0	74.0	42.0
Speaking	**	84.0	**
Writing	**	**	40.0
TotalScaleScore	92.0	92.0	84.0
PercentileRank	100.0	100.0	46.0

Clustered Instances
0 104 (76%)
1 32 (24%)

Fig 6-Percentile Score for UG Students Cluster Instances

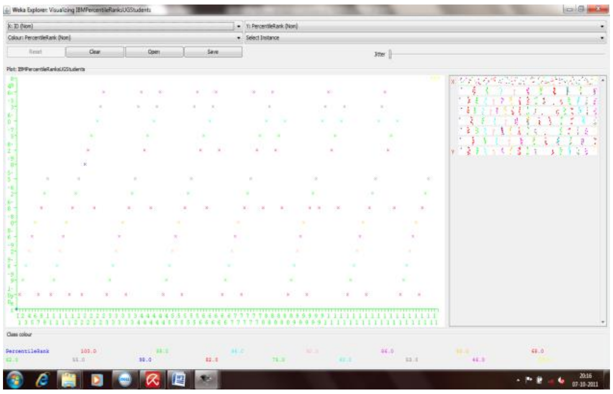

Fig 7-Percentile Rank for UG Students Graph(X&Y) Plot Axis Cluster Instances

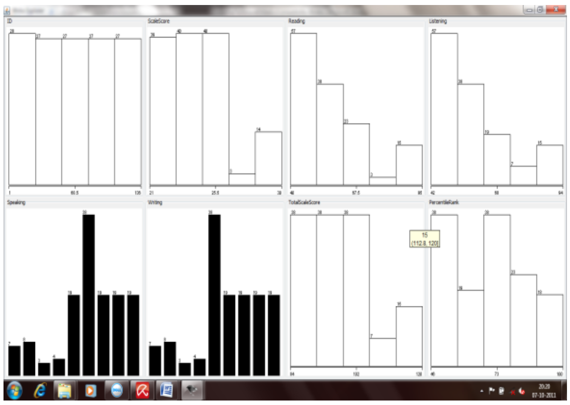

Fig 8-Percentile Rank for PG Students Visualize Attributes

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

```
Number of iterations: 4
Within cluster sum of squared errors:
213.10471161947248
Missing values globally replaced with
mean/mode

Cluster centroids:
                         Cluster#
Attribute          Full Data        0          1
                     (137)       (19)      (118)
=================================
================
ID                    68.5    26.1053    75.3263
Scale Score        24.2279    28.7895    23.4935
Reading            58.2647    87.5263    53.5531
Listening          59.7206    87.7895     55.201
Speaking                **       97.0         **
Writing                 **       95.0         **
TotalScaleScore            97.4412       116.4211
94.3851
Percentile Rank      69.25    97.6316    64.6801

                         Clustered Instances

                         0      19 ( 14%)
                         1     118 ( 86%)
```

Fig 9-Percentile Score for PG Students Cluster Instances

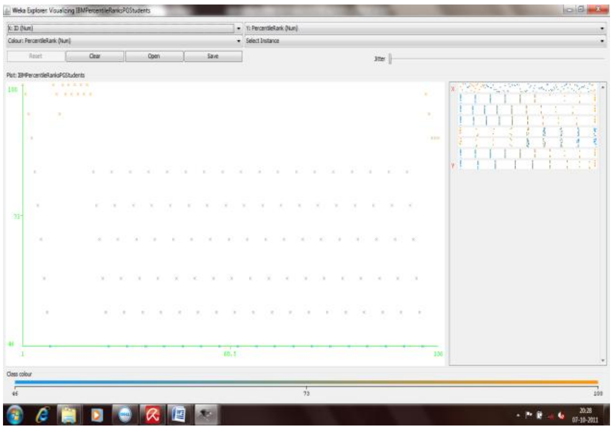

Fig 10-Percentile Rank for PG

Students Graph(X&Y) Plot Axis Cluster Instances

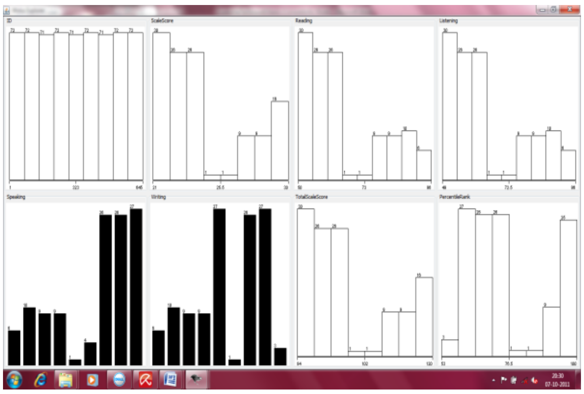

Fig 11-Percentile Rank for Other Students Visualize Attributes

```
Number of iterations: 3
Within cluster sum of squared errors:
227.63663687578602
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute          Full Data        0          1
                     (645)       (34)      (611)
===================================
=============
ID                     323    44.3235   338.5074
Scale Score        24.5339    28.3824    24.3197
Reading            65.9322    84.9118    64.8761
Listening          65.5424    85.2059    64.4482
      Speaking        59.0       98.0       59.0
      Writing           **       97.0         **
TotalScaleScore    98.1356   113.5294     97.279
Percentile Rank     74.161    96.2647     72.931

Clustered Instances
0      34 (  5%)
1     611 ( 95%)
```

Fig 12-Percentile Rank for Other Students Instances

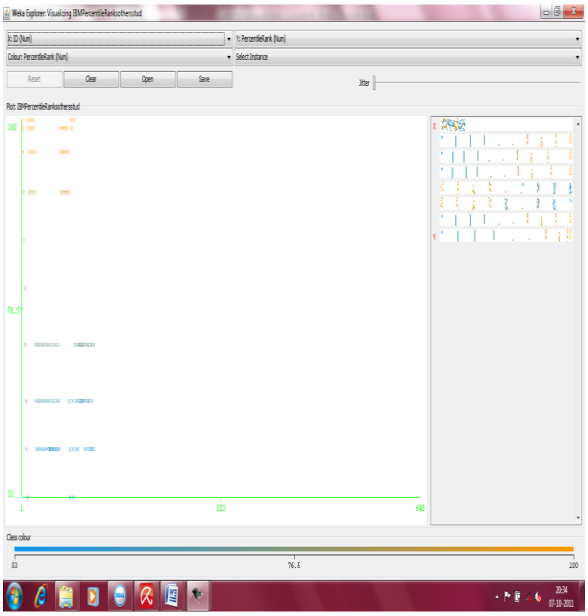

Fig 13-Percentile Rank for PG Students Graph(X&Y) Plot Axis Cluster Instances

VII. COMPARISON RESULTS

First of all, the differences between the different percentiles are very significant. For example, In this case consider Fig-4 shows the Percentile Score with the Cluster instances about UG/PG/Other Students Result success Rate in TOFEL Examination

Table-6

Percentile Scores/Instances (0+1=100%)	0	1	SSE
Percentile Scores(All)	11%	89%	209.65
Percentile Scores(UG)	76%	24%	863.00
Percentile Scores(PG)	14%	86%	213.10
Percentile Scores(Others)	5%	95%	227.63

In this comparison table shows the Percentile Scores for All Cluster 0 shows 11% ,but 1 shows 89% While Comparing UG Scores 0 Shows 76% 1 Shows 24% it gives much better result. In the SSE also is high in this category. In the third row PG and Others are 14%,5% respectively for 0 cluster,1 cluster show 86%,95% respectively. The findings that indicate that the more number of UG Students are scores are comparatively good than other Students.

VIII. CONCLUSIONS

In this study, K-Means Technique is used for finding the analysis of TOFEL IBT Score Comparison between the UG/PG/Other Students Percentile Score and SSE (Sum of Squared Error). This paper reports the innovative research that applies data mining technique to the data collected from the online examination portal. Recently there are many applications of data mining techniques carried out industrial as well as business domains.

At the centre of research this is real time analysis survey and the relationship will give the future directions and the better result.

FUTURE WORK

In future the data will be correlated with demographic mode in the future directions. How many people are qualifying in the sub continent especially in India? This research transforms the different kind of setup and give the proper knowledge to ramp up the preparation in the future. This paper will give the new transition to the student's community in future.

ACKNOWLEDGMENT

This research was supported by Dr.N.Kannan, Principal, Jayaram College of engineering and technology.

This research was also supported by Dr.A.Sahaya Arul Mary, Dean, Jayaram College of engineering and technology.

This research was also supported by Mr.A.Venkata Subramanian, HOD, IT Dept, Jayaram College of engineering and technology.

REFERENCES

1. Berry, Michael J. A. and Gordon Linoff, (1997), Data Mining Techniques, New York: John Wiley & Sons, Inc. .
2. [2] Mathematical Statistics with Applications. North Scituate, Massachusetts.
3. Introduction to Data Mining, Vipin Kumar,Ping-Pang,Michael Stein bench
4. www.ets.org
5. K. Pakhira, Malay, “A Modified k-means Algorithm to Avoid Empty Clusters,” International Journal of Recent Trends in Engineering, Vol 1,No. 1, May 2009.
6. R. Salman, V. Kecman, Q. Li, R. Strack and E. Test, “Two-Stage Clustering with k-means Algorithm,” WIMO 2011 Conference, Ankara,Turky, June 2011 (in press).

AUTHORS PROFILE

Mr. J. Emmanual Robin is working as Asst Professor in the Department of Computer Applications, Jayaram College of Engineering and Technology, Tiruchirappalli, Tamil Nadu, India. He has 2.5 years of experience in teaching and 6 years of experience in industry. He has presented 2 International / 3 National Conferences. His research interests are: Data mining and Web Services

Mr. G.Prabu is working as a Associate Professor in the Department of Computer Science and Engineering, Jayaram College of Engineering and Technology, Tiruchirappalli, Tamil Nadu, India. He has 1.5 years of experience in teaching and 15 years of experience in industry. His research interests are: Image Processing