Assessment of the Average Level of TOEFL Score by using SOM (Self Organizing Map) and K-Mean Clustering Techniques

Sekta Lonir Oscarini Watibhakti, Sujiati Jepriani, Bedi Suprapty, Rheo Malani

Abstract—Economic growth as measured by GDP growth rates and economic growth set as an increase in GDP strongly helps government predictions about the economic situation and the formation of economic development strategies. This measurement is done by combining mathematical and computer technology to make qualitative and quantitative predictions scientifically and appropriately for economic growth trends. It is a good practical sense to use scientific and proven methods to predict future GDP development trends of a particular economy. In some cases, machine learning methods have proven to be better forecasting results than statistical methods. A Deep Neural Network (DNN) is one type of ANN (Artificial Neural network) architecture based on deep MLP (Multi Layer Perceptron), which uses Deep Learning training techniques. This study proposes the use of DNN to predict the percentage of GDP distribution at current prices by industry sector. In this case, the DNN used will have multiple outputs as many industry sectors. The aim of this study is how to predict for the next period with the smallest possible prediction errors by using DNN.

Keywords: prediction, GDP, ANN, MLP, DNN.

I. INTRODUCTION

There are two widely accepted English proficiency tests for English learners: TOEFL and IELTS. The TOEFL (Test Of English as a Foreign Language) is made by ETS (Educational Testing Service), an institution in the United States. The TOEFL test is usually required for admission requirements at virtually all universities in the United States, Canada, Europe and Australia. There are three types of TOEFL, namely IBT (internet-based test), PBT (paper based test), and CBT (computer-based test). Now with technological advances, IBT TOEFL is practically more widely used. Generally this test takes about three hours and is organized in 4 parts, namely the section: grammar structure and written expression, listening comprehension, reading comprehension, writing. CBT Model is a test that uses computer media. Assessment scores of this model are in the range of 216-677. PBT model is a test that uses paper as a test medium. Assessment score with this model is in the range of 450-550 and above. The TOEFL exam score ranges from: 310 (minimum value) to 677 (maximum value) for PBT (paper-based test), while the value for IBT is from 0 to 120.

IELTS (International English Language Testing System) is a test of English proficiency held jointly by Cambridge University, British Council and IDP Education Australia. In the IELTS test, English used is British. The Global Test of English Communication (GTEC) created by the BurosCenter for Testing (Buros) is one tool that assesses English proficiency in reference to TOEFL and IELTS. Since 2017, GTEC test results can be received at 147 Japanese universities and 8 in the US [1].

The concept of an integrated model in teaching the language skills is necessary both in concept and in its instructions. The concept of an integrated model is an integrated skill that views the skills of listening, speaking, reading, and writing as an integral and inseparable part, as a conjunctional skill to be practiced in the classroom so that teaching and learning are truly communicative and productive. It is almost impossible to teach a second language and / or a foreign language without reading, speaking, and listening skills. Writing skills are also suitable for integrating skills for both teaching and assessment purposes. Reading is also considered an integral part of writing exercises in reader response theory, writing for reading, and reading for writing. In practice, there are many different types of language skills teaching models: source-based writing tasks, source-based reading tasks, source-based-listening tasks, source-based speaking tasks, and a mix of them. There has been so much research done to test the success rate among the various models as in [2] which has compared the results of language skills through writing independent .vs. integration essay. Several studies have also linked language proficiency with GPA (Grade Point Average) as in [3].

The success of teaching language skills is not only seen from how many students have passed the language test. Furthermore, it is also necessary to consider other factors that support the improvement of language skills through the evaluation of the results of some of the test of language skills that have been done. In this case historical data test results are required. There are many methods to test the success rate of teaching language skills through historical data of TOEFL test results. One of the most practical methods is to calculate

Revised Version Manuscript Received on 10 September, 2019.

SektaLonirOscariniWatibhakti, Department of Civil Engineering, State Polytechnic of Samarinda, East Kalimantan, Indonesia.(Email: sektalonir1226@gmail.com)

SujiatiJepriani,Department of Civil Engineering, State Polytechnic of Samarinda, East Kalimantan, Indonesia.(Email: sujiati_jepriani@polnes.ac.id)

BediSuprapty,Department of Information Technology, State Polytechnic of Samarinda, East Kalimantan, Indonesia.(Email: bedirheody@gmail.com)

RheoMalani, Department of Information Technology, State Polytechnic of Samarinda, East Kalimantan, Indonesia.(Email: anaogie@gmail.com)

the average value of student achievement during the TOEFL test. However, the average score does not reflect the overall performance of the progress. The most practical and significant way is to use clustering techniques. All data will be clustered into multiple clusters based on similarity level of data. Each cluster will have the centroid as the average value of all its cluster members which also reflects the characteristics of all its members.

There are many methods that can be used for clustering, both statistical methods and machine learning-based methods. K-Mean clustering is the most primitive, simple and effective method. This method has been used to mining the attitude of social network users and earthquake epicenter clustering in [4, 5].The only drawback is in the inconsistency of members of each cluster when new centroid initialization is used. Research on improving K-Mean ability has been done in [6]. Hierarchical clustering is one of the statistical methods, one of which has been used to classify stock market [7]. Fuzzy Logic concept has also been used to improve various clustering methods as described in several studies in [8-10]. Weighted Clustering algorithm (WCA) is one of the proposed clustering methods, based on combined weight metrics including node level, distance associated with neighboring node, node velocity, and time spent as cluster head [11]. Another technique for clustering is interactive clustering. This algorithm starts with an initial clustering and only make local changes in each step [12]. The model-based clustering assumes that a dataset to be clustered consists of various clusters with different distributions. The entire dataset is modeled by a mixture of these distributions. One study has applied this method where AHP (Analytical Hierarchy Process) is used to determine the effective number of clusters [13]. SOM (Self-Organizing Maps) is one of Artificial Neural Network architecture which is used for clustering that works with one type of machine learning that is unsupervised learning. Some research on SOM has been done in [14, 15].

This study applies SOM and K-Mean clustering to categorize historical data of students' TOEFL score. Historical data of students' TOEFL score containing TOEFL test results from the same students used as sample data in this study. The purpose of this study is to assess the success rate of teaching English proficiency based on clustering results.

II. METHODS

K-Mean

K-mean clustering belongs to a hard partitioning algorithm, which divides the multi-attribute data set into a number of K clusters [5]. Each individual in the dataset is allocated entirely to a particular cluster based on the closest distance to its centroid. First, a number of K centroids are initialized in a particular way in which each pair of centroid attributes must be chosen such that it is within the range of the dataset. The distance between each individual in the dataset against each centroid is calculated. Each individual will be considered a member of a cluster that has the shortest distance to its centroid cluster. The new centroids is generated based on the average attribute value of all its cluster members. The process is repeated until there is no change in cluster membership. K-Mean algorithm can be described as follows:

1. Read dataset (P). If the data set has a number of N data and M attributes, then data set is declared with

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,M} \\ p_{2,1} & p_{2,2} & \dots & p_{2,M} \\ \dots & \dots & \dots & \dots \\ p_{N,1} & p_{N,2} & \dots & p_{N,M} \end{bmatrix} \quad (1)$$

2. Determine the number of clusters (K).

3. Initialize a number of K centroids (C). Because the data set has a number of M attributes, the initial centroid is declared with

$$C = \begin{bmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,M} \\ c_{2,1} & c_{2,2} & \dots & c_{2,M} \\ \dots & \dots & \dots & \dots \\ c_{K,1} & c_{K,2} & \dots & c_{K,M} \end{bmatrix} \quad (2)$$

4. Calculate the distance between data with each centroid using Euclidean distance expressed as

$$d_{i,j} = \sqrt{\left(p_{i,1} - c_{j,1}\right)^2 + \left(p_{i,2} - c_{j,2}\right)^2 + \dots + \left(p_{i,M} - c_{j,M}\right)^2} \quad (3)$$

Where $d_{i,j}$ is distance between ith data and $j$thcentroid.

5. Grouping data based on the shortest distance against a particular centroid. If D is the distance matrix between data with each centroid then

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,N} \\ d_{2,1} & d_{2,2} & \dots & d_{2,N} \\ \dots & \dots & \dots & \dots \\ d_{K,1} & d_{K,2} & \dots & d_{K,N} \end{bmatrix} \quad (4)$$

The shortest distance for each data against each centroid is expressed by

$$D(i)_{min} = \begin{bmatrix} d_{1,i} \\ d_{2,i} \\ \dots \\ d_{K,i} \end{bmatrix} (5)$$

6. Calculate new centroid based on average of cluster members. For example, if the first cluster member is $P(1)$, $P(3)$, and $P(4)$ then

$$c_{1,1} = \frac{p_{1,1}+p_{3,1}+p_{4,1}}{3}, c_{1,2} = \frac{p_{1,2}+p_{3,2}+p_{4,2}}{3}, \dots, c_{1,M} = \frac{p_{1,M}+p_{3,M}+p_{4,M}}{3} \quad (6)$$

7. Repeat from step 4 until no more cluster membership changes.

Typically, centroids initialization is generated randomly within the range of the dataset. To ensure consistency of clustering results, in this study used the scattered average technique for its centroids initialization expressed by

$$P_{min} = [p_{min,1} \quad p_{min,2} \quad \dots \quad p_{min,M}]$$
$$P_{max} = [p_{max,1} \quad p_{max,2} \quad \dots \quad p_{max,M}]$$
$$part = (P_{max} - P_{min})/K \quad (7)$$
$$C(i) = P_{min} + (i-1)*part$$

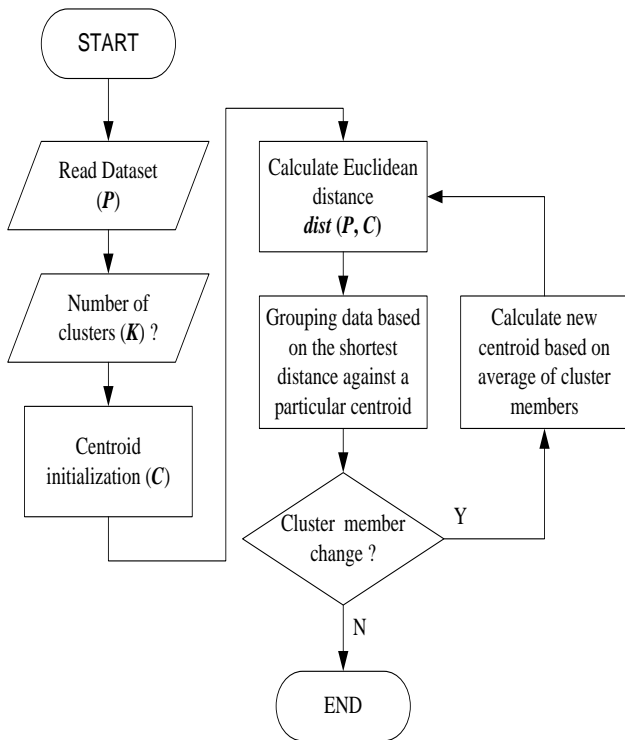Generally, K-Mean algorithm is shown in Fig.1.

Figure 1. K-Mean algorithm

SOM (Self Organizing Map)

The SOM algorithm is used to identify the natural patterns of the data set. The goal of SOM is to map the input data patterns into the n-dimensional grid of neurons. The grid forms an output space based on its neuron function. Due to the non-linear nature of its neuron function, SOM is capable of having more complex grouping geometries [15]. Another advantage of SOM is the organization of group maps following patterns in the output space that correlate with changes in input features that are most responsible for the result of clustering. SOM is one type of NN belonging to unsupervised learning. SOM architecture consists of input layer with *M* training vector unit, output layer with *K* cluster unit and intra-layer connecting between input layer and output layer. Each neuron in the input layer is directly connected to each neuron at the input layer where each relationship has a weighted vector of length *M*. As in K-Mean that has been described, *M* is the number of data attributes. SOM architecture is shown in Fig. 2.
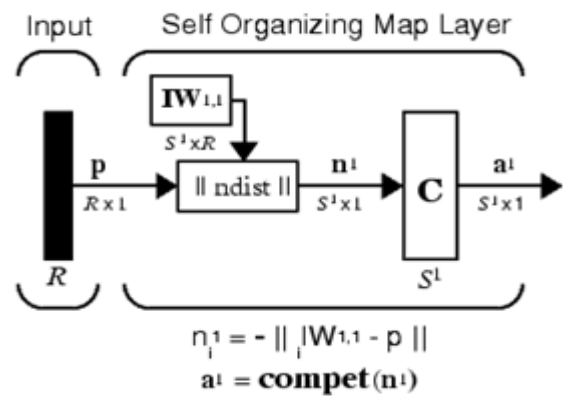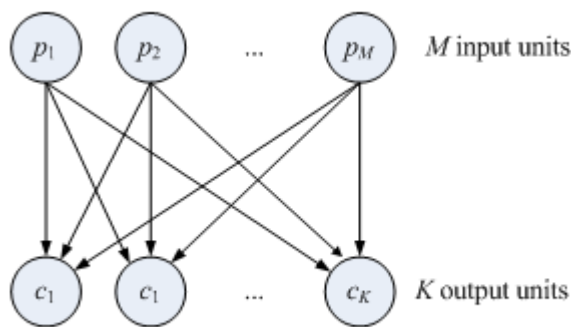

Figure2.SOM architecture

Intra-layer in principle is a function of neighborhood distance between the data at the input layer. This function will later be mapped between the data on the input layer into the cluster membership in the output layer. The final weighted result of the intra-layer acts as the centroid of each cluster. The neighboring distance is the distance between a weighted layer against each data expressed by

$$D = ndist(P, IW) = \|IW - P\|_2 \tag{8}$$

P is dataset, *IW* is weighted layer matrix, and *D* is distance matrix. The distance function used is Euclidean distance.

The winning neuron is the output neuron weights IW_i that have the shortest distance to *P*, expressed as

$$C(p) = \arg\min\|Iw_i - p\|_2 \quad i = 1\ldots K \tag{9}$$

Layer weighting initialization is also done with the scattered average technique. Layer-weighted updates are performed if there is still a change in cluster membership, expressed by

$$Iw_j(t+1) = Iw_j(t) - \eta\left(p_i - Iw_j(t)\right) \tag{10}$$

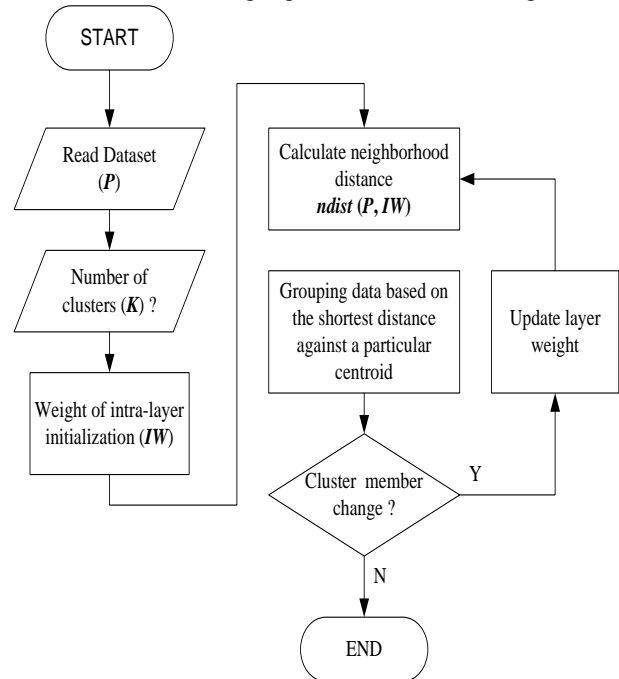η is learning rate.

SOM's training algorithm is shown in Fig. 3

Figure3.SOM's training algorithm

The optimum global solution

To achieve the optimum global solution, the distance of all cluster members to its centroid must be minimal and the distance between centroids should be maximally possible. If a data is known as follows:

$$X = [x_1 \quad x_2 \quad ... \quad x_n]$$

where n is the number of data, then the measure of the data distribution to the average of data is expressed by:

$$Var[X] = E[(X - \mu_x)^2] = \frac{1}{n}\sum_{i=1}^{n}(X - \mu_x)^2 \qquad (11)$$

Where $Var[X]$ is the variance of X and μ_x is the average of X.

If $Var[member]$ denotes the variance of the distance of all cluster members to its centroid, and $Var[centroid]$ denotes the variance of the centroid distance then the optimum global solution can be achieved if it satisfies:

$$F = min(Var[member]) \text{ and } max(Var[centroid]) (12)$$

Performance of optimum global solution is then expressed by:

$$P = \frac{Var[centroid]}{Var[member]} \qquad (13)$$

Dataset of students' TOEFL score

In this study is using historical data of students' TOEFL score from one department in State Polytechnic of Samarinda. As sample data is taken a number of 100 students from the first semester. The sample data is shown in Table 1. Data is taken from period 2016-2017. Percent progress is calculated using the following formula:

$$progress = \frac{(score\ 2017 - score\ 2016)}{score\ 2016} \times 100 \quad (14)$$

Table 1. Datasets of Students' TOEFL Score Period 2016-2017

No	2016	2017	Progress	No	2016	2017	Progress	No	2016	2017	Progress	No	2016	2017	Progress
1	487	507	4.11	26	493	487	(1.22)	51	320	380	18.75	76	380	357	(6.05)
2	477	453	(5.03)	27	473	430	(9.09)	52	320	363	13.44	77	390	340	(12.82)
3	477	440	(7.76)	28	470	393	(16.38)	53	323	401	24.15	78	434	447	3.00
4	473	467	(1.27)	29	457	433	(5.25)	54	434	450	3.69	79	393	397	1.02
5	473	403	(14.80)	30	447	417	(6.71)	55	270	370	37.04	80	337	361	7.12
6	473	403	(14.80)	31	443	397	(10.38)	56	377	393	4.24	81	424	383	(9.67)
7	473	410	(13.32)	32	443	450	1.58	57	395	417	5.57	82	394	399	1.27
8	457	390	(14.66)	33	440	430	(2.27)	58	419	384	(8.35)	83	453	443	(2.21)
9	457	413	(9.63)	34	430	413	(3.95)	59	377	487	29.18	84	387	356	(8.01)
10	450	500	11.11	35	427	343	(19.67)	60	323	411	27.24	85	402	404	0.50
11	443	440	(0.68)	36	423	357	(15.60)	61	403	313	(22.33)	86	331	330	(0.30)
12	433	420	(3.00)	37	413	393	(4.84)	62	336	397	18.15	87	337	367	8.90
13	430	367	(14.65)	38	410	413	0.73	63	394	316	(19.80)	88	354	386	9.04
14	410	380	(7.32)	39	397	373	(6.05)	64	341	359	5.28	89	393	317	(19.34)
15	403	360	(10.67)	40	397	367	(7.56)	65	356	339	(4.78)	90	394	407	3.30
16	397	443	11.59	41	393	393	-	66	337	419	24.33	91	319	341	6.90
17	390	407	4.36	42	393	380	(3.31)	67	356	361	1.40	92	357	446	24.93
18	387	400	3.36	43	383	453	18.28	68	323	304	(5.88)	93	317	407	28.39
19	380	410	7.89	44	380	373	(1.84)	69	324	337	4.01	94	381	376	(1.31)
20	380	403	6.05	45	380	367	(3.42)	70	357	376	5.32	95	362	353	(2.49)
21	373	407	9.12	46	370	377	1.89	71	347	320	(7.78)	96	353	347	(1.70)
22	363	367	1.10	47	370	387	4.59	72	378	354	(6.35)	97	361	316	(12.47)
23	357	367	2.80	48	363	370	1.93	73	351	287	(18.23)	98	432	370	(14.35)
24	357	357	-	49	353	323	(8.50)	74	326	403	23.62	99	342	430	25.73
25	313	363	15.97	50	353	347	(1.70)	75	460	334	(27.39)	100	336	316	(5.95)

III. RESULT AND DISCUSSION

Referring Table 1, there are three attributes of data that can be used i.e. TOEFL score in 2016 (X), TOEFL score in 2017 (Y), and progress in percent (Z). The TOEFL score is grouped into 3 clusters defined as *Low*, *Medium*, and *High*. The TOEFL score clustering based on pair of X and Y attributes is used to assess the final result, whereas the TOEFL score clustering based on pair of Y and Z attributes is used to assess the final progress, and finally, the TOEFL score clustering based on pair of X and Z attributes is used for the validation of the final result. The clustering results for each pair are shown in Fig. 4, 5, and 6, respectively, whereas final centroids and its members are shown in Table 2, 3, and 4, respectively.
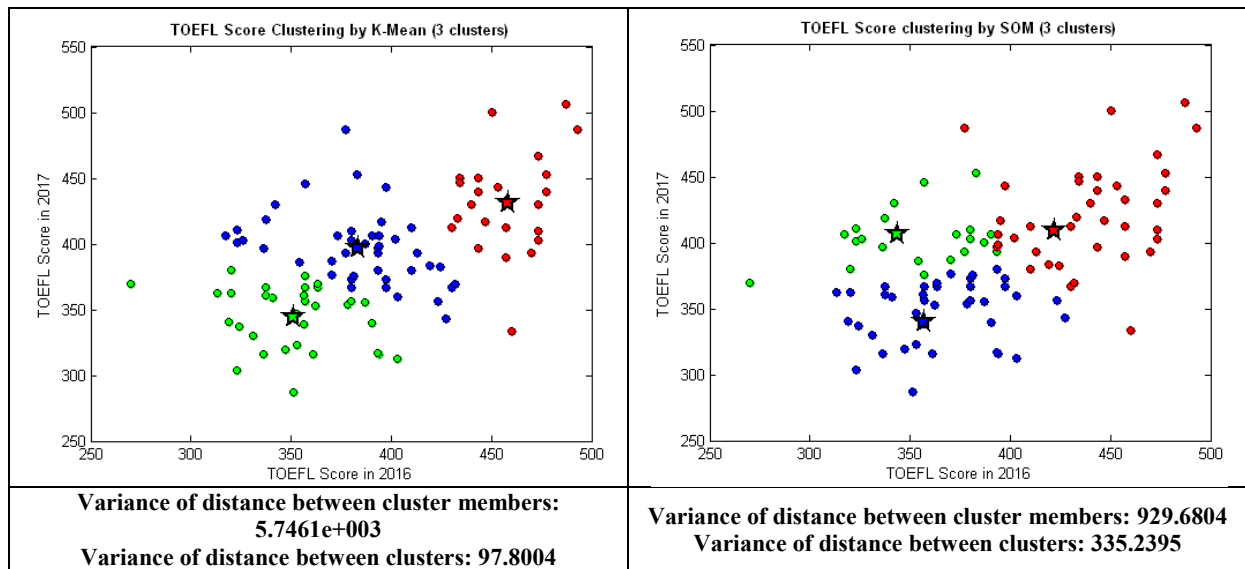
Figure4.Clustering results based on pair of *X* and *Y* attributes

Table 2. Final centroids and members from clustering results based on pair of *X* and *Y* attributes

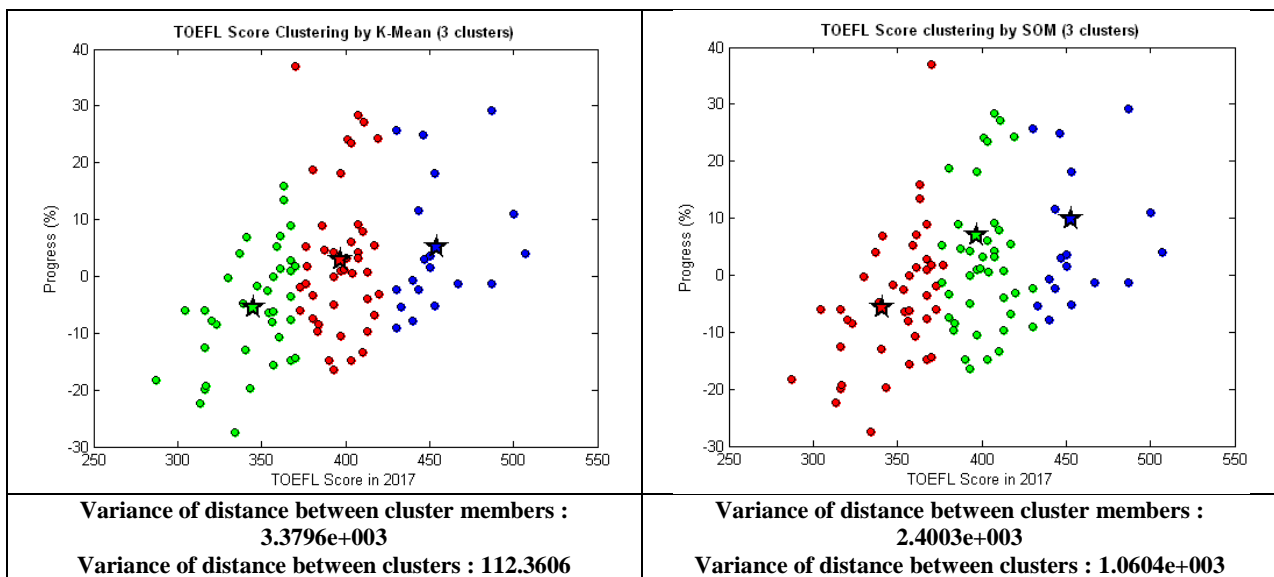| Cluster	K-Mean			SOM		
	X	Y	Members	X	Y	Members
1	442.4194	435.7097	31	423.7927	435.5585	31
2	389.5556	360.0926	54	385.7179	350.8480	55
3	322.9333	385.1333	15	321.3624	390.1833	14

Figure5.Clustering results based on pair of *Y* and *Z* attribute

Table 3. Final centroids and members from clustering results based on pair of *Y* and *Z* attributes

Cluster	K-Mean			SOM		
	Y	Z	Members	Y	Z	Members
1	396.8605	2.7815	43	340.6402	-5.6791	42
2	345.0789	-5.4983	38	396.7091	6.9022	41
3	454.5263	5.1794	19	452.7455	9.8855	17

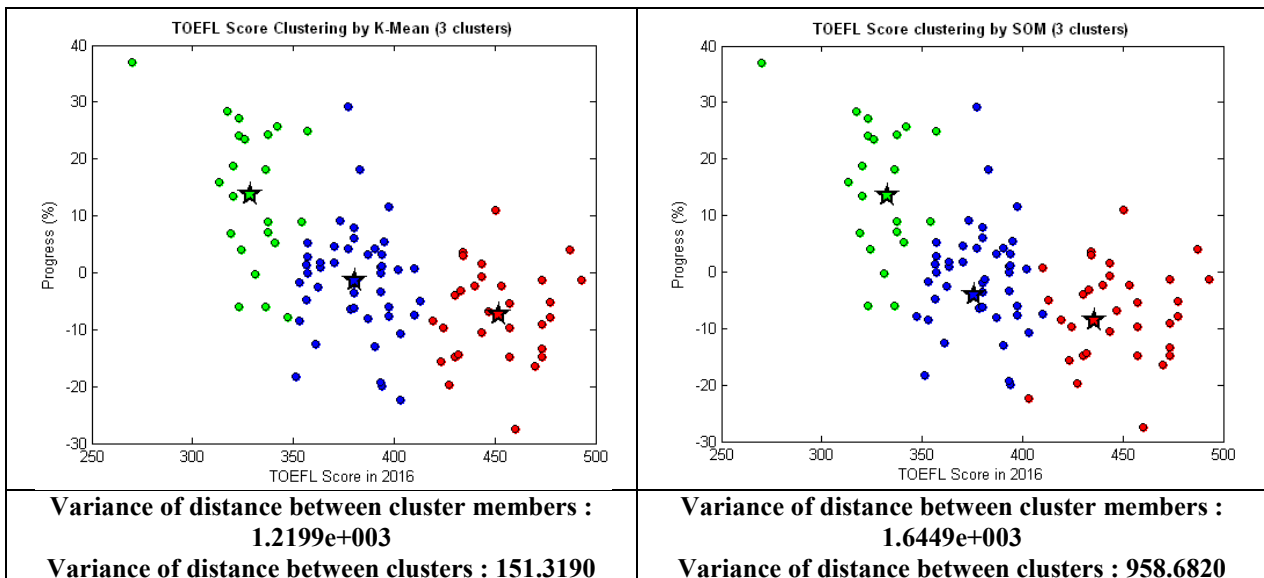| Variance of distance between cluster members : 1.2199e+003	Variance of distance between cluster members : 1.6449e+003
Variance of distance between clusters : 151.3190	Variance of distance between clusters : 958.6820

Figure6.Clustering results based on pair of *X* and *Z* attributes

Table 4. Final centroids and members from clustering results based on pair of *X* and *Z* attributes

Cluster	K-Mean			SOM		
	X	Z	Members	X	Z	Members
1	451.7742	-7.3750	31	435.9438	-8.6171	34
2	328.7727	13.7765	22	332.4411	13.4708	21
3	380.5532	-1.3910	47	375.6442	-4.1203	45

3.1 Variance of distance analysis

Referring Fig. 4, 5, and 6 then obtained the variance of distance analysis using performance of optimum global solution as in Eq. (13) with the results as shown in Table 5.

Table 5. Variance of distance analysis

Pair	K-Mean			SOM		
	between clusters	between cluster members	*P*	between clusters	between cluster members	*P*
X-Y	97.8004	5.7461e+003	0.0170	335.2395	929.6804	0.3606
Y-Z	112.3606	3.3796e+003	0.0332	1.0604e+003	2.4003e+003	0.4418
X-Z	151.3190	1.2199e+003	0.1240	958.6820	1.6449e+003	0.5828

Referring to Table 5 it is found that SOM clustering is better than K-Mean clustering. This is proven by SOM clustering having a larger P value compared to K-Mean clustering.

Since the end centroid reflects Low, Medium, and High performance levels it needs to be sorted ascending for the second attribute of each pair as shown in Table 6.

3.2 Analysis of consistency of the results

Table 6. Analysis of consistency of the results

clustering results based on the pair of *X* and *Y* attributes			
Cluster	*X* (2016)	*Y* (2017)	Members	progress
Low	385.7179	350.8480	55	(9.04)
Medium	321.3624	390.1833	14	21.42
High	423.7927	435.5585	31	2.78

Cluster	clustering results based on the pair of X and Z attributes			clustering results based on the pair of Y and Z attributes			
	X (2016)	Z (progress)	Members	Y (2017)	Z (progress)	Members	Difference
Low	435.9438	-8.6171	34	340.6402	-5.6791	42	8
Medium	375.6442	-4.1203	45	396.7091	6.9022	41	-4
High	332.4411	13.4708	21	452.7455	9.8855	17	-4

Referring to Table 6, from the clustering results based on the pair of X and Y attributes found that the highest cluster has an average TOEFL score of 424 in 2016 and increased to 436 in 2017 without considering the progress factor of each student. In this case, there was a progress of 2.78%. Based on the TOEFL PBT model which has an assessment score of 450-550, the average TOEFL score for the highest cluster of 2017 has not reached the minimum score.

From the clustering results based on the pair of X and Z attributes, it was found that the highest cluster in 2016, a total of 21 students with an average TOEFL score of 332 (X) achieved a 13.47% progression in 2017 (Z). But from the clustering results based on the pair of Y and Z attributes, it was found that only 17 students were able to achieve an average TOEFL score of 453 (Y) in 2017 with a 9.89% progress (Z).

It can be concluded that in general, without considering the progress factor of each student, the average TOEFL score in the highest cluster of 2017 has not been able to achieve the minimum requirement of TOEFL PBT model (450-550). With considering the progress factor of each student, only 17 students successfully fulfill the minimum requirement of TOEFL PBT model by achieving the average value of TOEFL 453 in 2017 with the progress of 9.89%

IV. CONCLUSION

In this study, SOM clustering and K-Mean clustering have been used to classify students' TOEFL scores and proved that SOM clustering is better than K-Mean clustering. Taking into account the progress factors of each student, the SOM clustering is able to significantly classify the TOEFL score of the students based on the progress achieved. In principle, clustering techniques applied to a student TOEFL score do not indicate the performance of each student, but each student in each cluster can be indicated for similarity in performance based on the centroid cluster. It can be concluded that the final centroid resulting from the application of clustering techniques on the TOEFL score dataset of students reflects the successful performance of teaching English proficiency. Future work is how to improve the significance of clustering results closer to the performance of each cluster member.

V. ACKNOWLEDGMENTS

REFERENCES

1. M. Kim, W. Z. Smith, and T.-Y. Chin. (2017). Validation and Linking Scores for the Global Test of English Communication: White Paper.
2. J. Gholami and M. Alinasab, "Source-Based Tasks in Writing Independent and Integrated Essays," International Journal of Instruction, vol. 10, pp. 127-142, 2017.
3. N. Yogendra and A. Andrew, "A Study On The Factors Influencing On G rad e Point Average (G PA) With Special Reference To Third Year Commerce And Management Students Of Eastern University, Sri Lanka," Journal for Studies in Management and Pl anning, vol. 3, 2017.
4. V. Gurusamy, S. Kannan, and J. R. Prabhu, "Mining the Attitude of Social Network Users using K-means Clustering," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, pp. 226-230, 2017.
5. P. Novianti, D. Setyorini, and U. Rafflesia, "K-Means cluster analysis in earthquake epicenter clustering," International Journal of Advances in Intelligent Informatics, vol. 3, p. 81, 2017.
6. A. Bansal, M. Sharma, and S. Goel, "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining," International Journal of Computer Applications, vol. 157, pp. 35-40, 2017.
7. A. Salighehdar, Y. Liu, D. Bozdog, and a. I. Florescu, "Cluster Analysis of Liquidity Measures in a Stock Market using High Frequency Data," Journal of Management Science and Business Intelligence, pp. 1-8, 2017.
8. B. S. Harish and S. V. A. Kumar, "Anomaly based Intrusion Detection using Modified Fuzzy Clustering," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4, p. 54, 2017.
9. J. Yang, Y.-s. Ke, and M.-z. Wang, "An adaptive clustering segmentation algorithm based on FCM," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 25, pp. 4533-4544, 2017.
10. Y. Zhang, J. Wang, D. Han, H. Wu, and R. Zhou, "Fuzzy-Logic Based Distributed Energy-Efficient Clustering Algorithm for Wireless Sensor Networks," Sensors, vol. 17, p. 1554, 2017.
11. M. Patil and R. C. Biradar, "Energy Efficient Weighted Clustering Algorithm in Wireless Sensor Networks," Global Journal of Computer Science and Technology: E-Network, Web & Security, vol. 17, pp. 54-63, 2017.
12. P. Awasthi, M. F. Balcan, and K. Voevodski, "Local algorithms for interactive clustering," Journal of Machine Learning Research, vol. 18, pp. 1-35, 2017.
13. S. Akogul and M. Erisoglu, "An Approach for Determining the Number of Clusters in a Model-Based Cluster Analysis," Entropy, vol. 19, p. 452, 2017.
14. A. Akande, A. C. Costa, J. Mateu, and R. Henriques, "Geospatial Analysis of Extreme Weather Events in Nigeria (1985–2015) Using Self-Organizing Maps," Advances in Meteorology, vol. 2017, pp. 1-11, 2017.
15. C. Nguyen, M. Starek, P. Tissot, and J. Gibeaut, "Unsupervised Clustering Method for Complexity Reduction of Terrestrial Lidar Data in Marshes," Remote Sensing, vol. 10, p. 133, 2018.

Retrieval Number: B13120982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1312.0982S1119

2599

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication