Malaya Journal of Matematik, Vol. 9, No. 1, 163-168, 2021

https://doi.org/10.26637/MJM0901/0027

Application of k-means clustering in spoken English training process

Kaja Mohaideen D[1]*, Baskaran S[2] and Mohammed Ibrahim D[3]

Abstract

Data Clustering is the most important data mining technique playing a vital role in various fields such as Business, Medicine, Construction, etc. In this study, k-means clustering technique is utilized to understand the skill level of the students enrolled in Spoken English Training (SET) programme and effectively strategize the need-based training sessions for them according to their present knowledge and requirements. Pre-training data collected from 159 students enrolled for Spoken English Training (SET) programmes in Chennai, India which consists of marks secured by the students in three tests concerning five basic categories namely content, communication, pronunciation, vocabulary, and grammar. All the necessary skills required in each category are thoroughly examined and scored. Data is then clustered using Elbow method and clustering technique to categorize the student mass into four different groups. The strengths and weaknesses of each group are uniquely diagnosed and necessary tailor-made curriculum and training sessions are advised so that effective suitable training can be given to each candidate at optimal time duration and cost.

Keywords

data mining, k-means clustering, Spoken English training process.

AMS Subject Classification

91C20, 62H30, 68T10.

[1,2] *PG Department of Mathematics, The New College, Chennai-600014, Tamil Nadu, India.*
[3] *Opportunities Infinite Training Academy, Chennai-600112, Tamil Nadu, India.*
***Corresponding author**: [1] kajamohaideen@thenewcollege.edu.in; [2]baskaran@thenewcollege.edu.in

Article History: Received **08** November **2020**; Accepted **19** December **2020**

Contents

1. Introduction

Analyzing and understanding big data is the biggest challenge in many present scenarios especially retrieving the crucial hidden matter and the underlying pattern from the data is the toughest task in many situations. Clustering analysis is an essential tool that is helpful in data analysis and in categorizing them into clusters of homogeneous elements [1]. Data matter processed under clustering analysis is categorized in many clusters such that the data particles present in a cluster with high similarity whereas those with variations are placed in different clusters. Various fields possess wide range of applications of data clustering processes such as image segmentation [2,3], text mining [4,5], bioinformatics [6,7], wireless sensor networks [8,9], and financial analysis [10]. There are many clustering methods and algorithms that are present in the literature. Among them k – means clustering is simple, efficient, and highly popular clustering method which has tremendous real-time applications [11]. In this study, the application of the k-means clustering technique in improving the English Language Training(ELT) Process is explored.

The third most communicative language in the world is English which is also considered as the official workplace language in the majority of industries such as information technology, engineering, media, education, politics, etc. During the past decade, the importance of uplifting fluency in English is drastically increased in developing counties which are ultimately resulted into the outsourcing of millions of jobs from developed countries to developing countries [12]. Especially in India, fluency in the English language plays a major role in getting a white-collar job with a decent salary. Enormous research took place in this field has established that fluency in the English language is related to 32% increased earnings in India [13]. The English language is the medium of the teaching-learning process and research in the majority of educational institutions in India. The ability of fluent and errorless spoken and written expertise in English is recognized as a key skill necessary for acquiring a managerial post and further promotions in many industries and fields in India. Concerning the pilot strategy in India, the nation's New Education Policy draft underscored the significance of showing English as a major aspect of the three-language equation in school education [14].

As the requirement of communication skills builds in many industries, teaching and testing the English as a foreign language has absorbed enormous attention in India. Especially, many researchers working on discovering innovative methodologies to effectively improve the quality of teaching and learning process of English. In particular, English language proficiency is one of the most important skills necessary to acquire a better job in India and worldwide. Lack of communication skills (spoken and written) becomes a major factor affecting graduates of various Indian universities for not getting their dream job. The potential expenses of obtaining quality communicative skills include Course fee, time, distance and efforts, etc. Due to poverty, many youngsters are not in the state to invest huge amount of money in communication training courses. Another major challenge is that many students possess erroneous conception about their skill level in English proficiency and finally struggle at the time of their job interviews.

Overcoming all these barriers, every year thousands of young graduates attending English language training programmes to enhance their communication skills but the results are not so impressive. Many participants don't acquire what they really need and which directly decreases their employability chances. A unique structured course idea to the heterogeneous students with dissimilar categories of English skill requirement is unwarranted which will be just a waste of time, effort, and resources. It becomes necessary to diagnose the skill area of each student, identify the shortcomings, and address them with the required amount of effort. In this study, the scores of Pre-Course assessment tests conducted on 159 candidates enrolled for spoken English training clustered using the Elbow method and k-means clustering technique to classify them into homogeneous groups with similar requirements so that their training sessions may be planned accordingly. Moreover, such scores avail opportunities to the participants to cross-check their self-estimation about their spoken English proficiency.

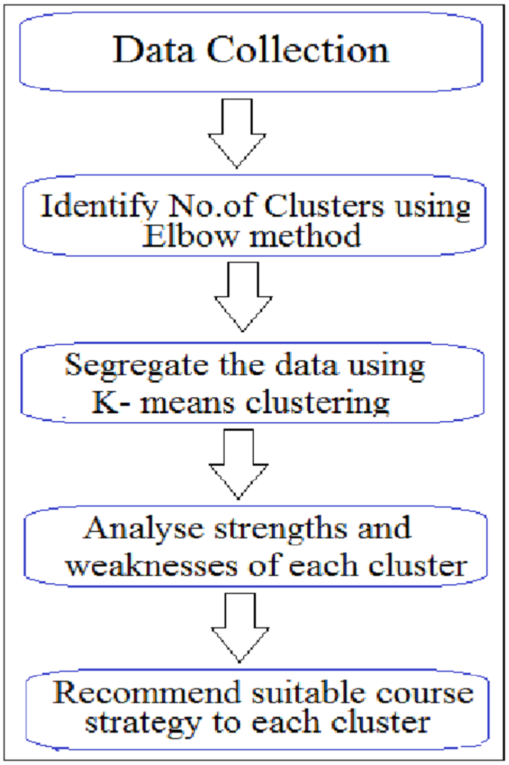
Data Collection
Identify No.of Clusters using Elbow method
Segregate the data using K- means clustering
Analyse strengths and weaknesses of each cluster
Recommend suitable course strategy to each cluster

Figure 1. Research Methodology

2. Methodology

The research methodology in this case study design can be identified in to five stages in which the data collection process is the first stage, Identifying the efficient number of clusters to be made out of collected data using Elbow method algorithm is the second stage. Data mining framework applying k means clustering algorithm is in third stage. In the fourth stage, the clusters acquired from data mining are utilized to categorize the students into various groups and analyzed to identify the underlying patterns in the spoken English skills of each group. The final stage proposes the strategy to be adopted in designing curriculum and class allocation for every individual cluster of participants according to their requirements (Figure 1).

2.1 k-means clustering method

k-means clustering is an unsupervised machine learning process that is efficiently employed for partitioning a specific dataset into k groups called clusters. It is a technique intended to distinguish concealed patterns from the data to foresee possible activities, expected trends, and group (cluster) data based on similarities [15,16]. In most of the cases, the Euclidean distance is defined as the similarity between a pair of objects. The number of clusters k is estimated by the expert to develop an efficient clustering of data. Each generated cluster is characterized by its centroid and is also termed as the mean value of the cluster.

The main process is to minimize the objective function, that is, the Sum of Squared Errors

$$H = \sum_{i=1}^{k} \sum_{j \in \mathscr{C}_i} ||x_j - \tau_i||^2 \tag{2.1}$$

where H is the objective function, x_j is the jth element, $\mathscr{C}_i$ and τ_i are data set and center of ith cluster respectively and k represents the total number of clusters. The aim of the process is to minimize the distance of each data particle x_j from the center τ_i contained in every cluster C_i [17].

The algorithm for k means clustering process is given below

Step 1: Fix the primary set of k centroids.

Step 2: Assign each and every element or data particle to the cluster with nearest centroid.

Step 3: Calculate the mean of each cluster and identify the most suitable k centroids.

Step 4: Repeat steps 2 and 3 until there is no change in the criterion function after an iteration [18].

2.2 Elbow Method

Elbow method is a visual process crucial to identify ‘k’, the number of clusters required for effective clustering of the data using k-means algorithm. It is known that the main process of k-means is grouping the given data in such a way that the total variation within a given cluster is minimum. The Elbow method observes The total within-cluster sum of squares (WSS) as a cluster number function appropriate for the dataset under process and resulting in to a suitable k, the number of clusters in such a way that adding more clusters does not have any significant impact on WSS and the data analysis results.

Elbow method Algorithm works as follows:

Step 1: For each k value, k-means clustering is computed.

Step 2: The total within-cluster sum of squares (WSS) is computed for each k.

Step 3: WSS curve is plotted against ‘k’.

Step 4: Elbow like bending in the graph is the real pointer of the optimal number of clusters.

3. Case Study

3.1 Data collection

Opportunities Infinite Training Academy (OITA) is a private firm established in the year 2007 in Chennai, Tamilnadu, India. They are pioneers in training the participants in developing spoken English proficiency and employability skills. Every year thousands of students from various parts of Tamilnadu enrolling in their course to improve their spoken English skills. The Firm also undertakes assignments in various Engineering Colleges for training their final semester outgoing BE/BTech students in personality development, employability skills, and spoken English proficiency. OITA’s spoken English course structure offers 100 hours of training sessions over six weeks. Each session lasts for two hours, and students are assessed at the end of the course. The medium of instruction is both English and Tamil which plays an essential role in smoothening the communication between students and teachers. The course primarily focuses on the skill set of spoken English required for a candidate to perform in a job interview confidently and complete it successfully. They also equip the candidate with the language expertise that is necessary in various middle and high profile industries focusing on meeting and connecting with clients, associates and customers. OITA adopts classroom instruction, peer discussion, and electronic media (to build vocabulary and improve phonetics) methods to train the candidates.

In this study, 159 students enrolled for summer vacation special courses conducted during (April 2018 to June 2018) were considered. At the time of admission, each student filled up an application form recording necessary details such as Name, Age, Sex, Address, Qualification, Contact, etc for registration which is summarized in Table 1. After the admission process, every student has to undergo a Pre-Course Spoken English Assessment Test in OITA. Totally 12 senior faculty members possessing more than 5 years of training experience in this field are invited to conduct the test. Six batches of two faculty members each are formed in which the first three batches (A, B, and C) are allotted with 27 students each and the next three batches (D, E, and F) are allotted with 26 students each respectively. The whole testing process took 5 hours (9 am to 2 pm) with a refreshment break of 15 minutes at 11.30 am.

The test consists of three parts namely Self Introduction in which the ability of candidates to express themselves are tested, they were expected to speak about their strength and weakness, their future ideas and targets, etc. second part Group Participation in which students are involved discussions and conversations, role-playing games, etc. Intensive Speaking is the third part of the test in which each student is given a print out consist of three pictures in which they have to select any one and prepare themselves to share their ideas about it for 2-3 minutes without any interruption. Other participants were prohibited to talk during their partner’s speech, they were made to listen carefully as the faculty members asked a few questions at the end of each speech. The participants who answered three questions were not given a chance for the next questions to ensure the involvement of each and every candidate. The test process ends up with a vote of thanks from a faculty member of each batch.

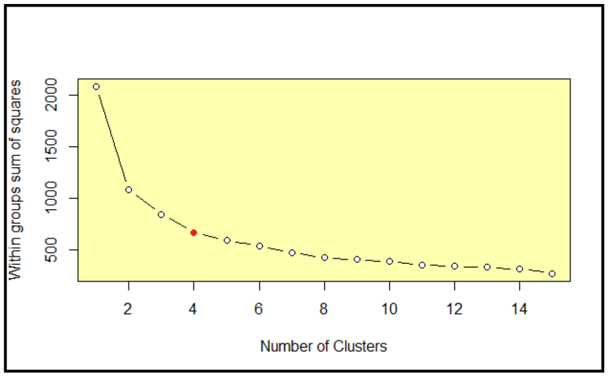

Figure 2. Estimation of the number of clusters using Elbow method

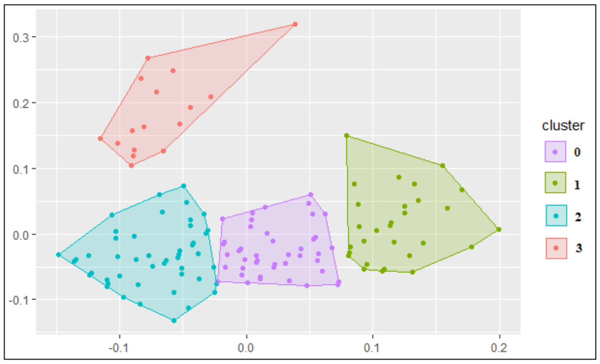

Figure 3. Clustering of data using R programming

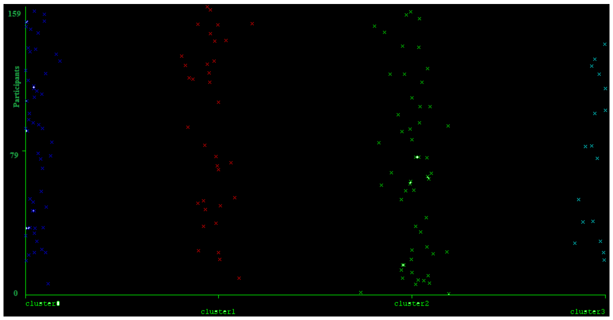

Figure 4. Clustering of data using WEKA

The five criteria namely content, communication, pronunciation, vocabulary, and grammar were evaluated for each candidate. Skill weight for each criterion is marked in 10 points Likert scale (0 to 10). A student with no skill in any criteria assigned with 0 and outstanding skill assigned with 10. The marks obtained by the candidates in each criterion are displayed in Table 2.

Table 1. Demographic data of participants

Gender		Age				Qualification(completed)		
Male	Female	< 18	18-24	24-28	> 28	HSC	UG	PG
66	93	22	52	58	27	32	97	30

Table 2. Format example of participant's dataset

Instance number	Reg.No	content	Communication	Pronunciation	Vocabulary	Grammar
1	18SET001	8	5	7	4	7
2	18SET002	8	6	9	5	7
3	18SET003	9	7	10	5	9
4	18SET004	7	4	7	2	6
5	18SET005	10	7	9	7	10
6	18SET006	6	7	10	7	6
7	18SET007	8	6	8	5	8
8	18SET008	9	7	10	6	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮
148	18SET148	5	2	5	1	5
149	18SET149	7	5	6	3	6
150	18SET150	8	4	6	2	7
151	18SET151	5	4	6	1	5
152	18SET152	8	6	9	4	8
153	18SET153	5	3	5	1	4
154	18SET154	10	6	8	4	8
155	18SET155	8	5	8	3	7
156	18SET156	9	6	8	4	8
157	18SET157	9	5	8	3	8
158	18SET158	6	5	6	1	6
159	18SET159	8	6	8	3	8

3.2 Data Analysis

3.2.1 Elbow method

The collected data of participants processed using Elbow method to identify the optimal number of clusters as 4 (ie n=4). It is found that there is no significant change occurs in WSS value if the number of clusters increased beyond n=4 (Figure 2).

3.2.2 k–means clustering of the data

Data segregated into four clusters using k-means algorithm. The data mining software WEKA 3.8 utilized for clustering the data into four clusters and also R programming software used to cross-check the obtained results. It is found that both software establishes the utmost($\approx 99\%$) same results (Figure 3 and 4) which are tabulated in Table 3.

Table 3. k-means clustering results

	Cluster Centroids				
Attribute	**Full Data**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Content	7.7799	7.6604	5.8529	8.9273	8.2941
Communication	4.9434	4.3962	3.0294	5.9091	7.3529
Pronunciation	6.9308	6.566	4.7059	8.2	8.4118
Vocabulary	4.2138	3.6226	2.1176	5.3455	6.5882
Grammar	7.6226	7.6981	6.0882	9.1273	5.5882
No.of Instances	159 (100%)	53 (33%)	34 (21%)	55 (35%)	17 (11%)

4. Results and Discussions

Results of clustering presented show plenty of variations in the skill set of participants of various clusters. On the whole, It can be seen that the skill level of participants in 'Content' and 'Grammar' is appreciable whereas that in 'communication' and 'Vocabulary' is very low. Category wise analysis for the skill level of various clusters is recorded below. The skill weight of 7.5 is considered as the minimum marks expected in each category.

4.1 Analysis of 'Content'

It can be inferred from Figure 5

Rigorous: The skill level of cluster1 is very low with

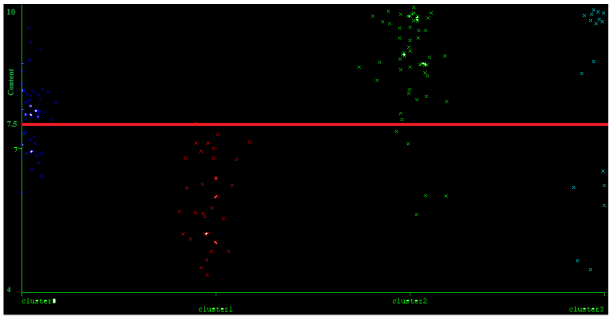

Figure 5. Clustering of 'Content' data

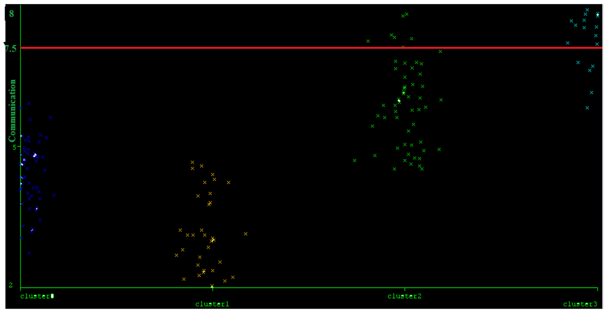

Figure 6. Clustering of 'Communication' data

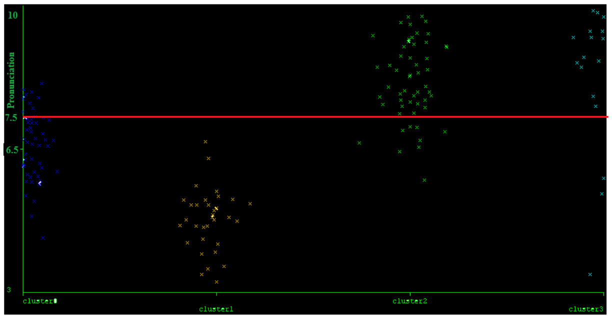

Figure 7. Clustering of 'Pronunciation' data

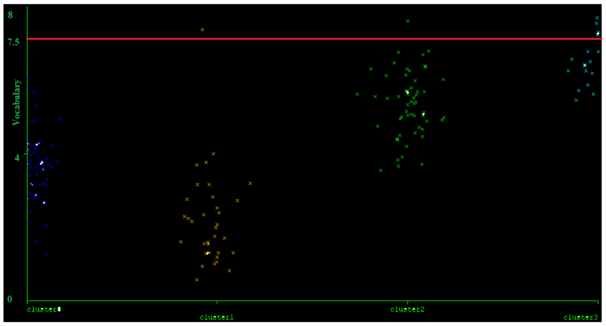

Figure 8. Clustering of 'Vocabulary' data

respect to the expected level. All the participants of cluster1 need rigorous training for developing their skills in 'content'.

Moderate: Some participants of cluster0 and Cluster3 are above the expected level and the majority of them near the threshold level hence they need moderate training.

Slight: majority of participants of cluster2 are above cut off level, and others are very close, and hence slight training is enough for them in this category.

4.2 Analysis of 'Communication'

It can be inferred from Figure 6

Rigorous: The skill level of cluster0 and cluster1 is very low with respect to the expected level. All the participants of these clusters need rigorous training for developing their communication skills.

Moderate: Some participants of cluster2 are above the expected level and the majority of them near the threshold level hence they need moderate training.

Slight: more than 60% of participants of cluster3 are above 7.5, and others are very close, and hence slight training is enough for them in this category.

4.3 Analysis of 'Pronunciation'

It can be inferred from Figure 7

Rigorous: The skill level of cluster1 is very low as none of them is near the cut off level and they need rigorous training for developing their 'Pronunciation' skills.

Moderate: Some participants of cluster0 are above the expected level and the majority of them near the threshold level hence the moderate training may be enough for them.

Slight: the majority of participants of cluster2 and cluster3 are above cut off level, and others are close, and hence slight training is enough for them in this category.

4.4 Analysis of 'Vocabulary'

It can be inferred from Figure 8

Rigorous: The skill level of cluster0 and cluster1 is very low as none of them is near the cut off level and they need rigorous training for developing their 'Vocabulary' skills.

Moderate: Few participants of cluster2 and cluster3 are above the expected level but the majority of them near the cutoff level hence the moderate training may be enough for them.

4.5 Analysis of 'Grammar'

It can be inferred from Figure 9

Rigorous: The skill level of cluster1 and cluster 3 is very low and they need rigorous training for developing their 'Grammar' skills.

Moderate: Some participants of cluster0 are above the expected level and the majority of them near the threshold level hence the moderate training may be enough for them.

Slight: the majority of participants of cluster2 are above cut off level, and others are close, and hence slight training is enough for them in this category.

5. Conclusion

All the above discussions can be consolidated into the table which shows that the participants in each cluster have a unique

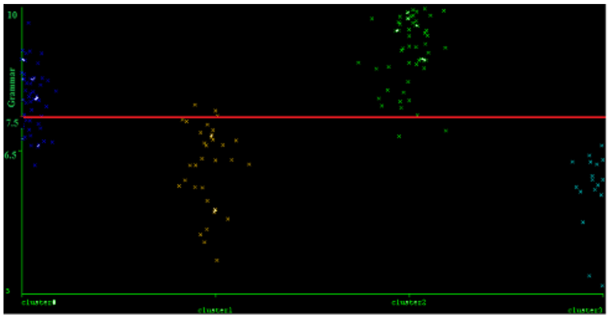

Figure 9. Clustering of 'Grammar' data

requirement of training which is completely different from all other clusters

Table 4. Cluster based skill analysis

	Content	Communication	Pronunciation	Vocabulary	Grammar
Cluster 0	Moderate	Rigorous	Moderate	Rigorous	Moderate
Cluster 1	Rigorous	Rigorous	Rigorous	Rigorous	Rigorous
Cluster 2	Slight	Moderate	Slight	Moderate	Slight
Cluster 3	Moderate	Slight	Slight	Moderate	Rigorous

It can be noticed from Table 4 that every cluster of participants requires a different curriculum structure that specifically focusing on their required category and a unique tailor-made training session fulfilling their shortcomings. Further, It can be seen that all the 34 participants of cluster1 require rigorous training to improve all the skill categories compare to 55 participants of cluster2 who require slight training in 3 categories (content, Pronunciation, Grammar) and moderate training in 2 categories (communication, vocabulary). It would be meaningless if the same curriculum design is followed to both clusters with the same 100 hours of training.

As course fees and the number of training hours is a major factor directly affecting admission of participants in SET courses it is very crucial to fix the number of training hours and curriculum suitable to the present skill level and requirement of the participant.

References

[1] Jain A. Data clustering: 50 years beyond K-means. Pattern Recognit Lett.31(8) (2010) 651–666.

[2] Gong M, Liang Y, Shi J, Ma W, Ma J. Fuzzy C-Means Clustering With Local Information and Kernel Metric for Image Segmentation. IEEE Transactions on Image Processing. 2013;22(2):573-584.

[3] Tu X, Gao J, Zhu C et al. MR image segmentation and bias field estimation based on coherent local intensity clustering with total variation regularization. Med Biol Eng Comput. 2016;54(12):1807-1818.

[4] Mahdavi M, Abolhassani H. Harmony K-means algorithm for document clustering. Data Min Knowl Discov. 2008;18(3):370-391.

[5] Chitra A, Rajkumar A. Paraphrase Extraction using fuzzy hierarchical clustering. Appl Soft Comput. 2015;34:426-437.

[6] Iván G, Grolmusz V. On dimension reduction of clustering results in structural bioinformatics. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2014;1844(12):2277-2283

[7] Triguero I, del Río S, López V, Bacardit J, Benítez J, Herrera F. ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem. Knowl Based Syst. 2015;87:69-79.

[8] Liu C, Lee C, Wang L. Distributed clustering algorithms for data-gathering in wireless mobile sensor networks. J Parallel Distrib Comput. 2007;67(11):1187-1200.

[9] Zhu J, Lung C, Srivastava V. A hybrid clustering technique using quantitative and qualitative data for wireless sensor networks. Ad Hoc Netw. 2015;25:38-53.

[10] Marinakis Y, Marinaki M, Doumpos M, Zopounidis C. Ant colony and particle swarm optimization for financial classification problems. Expert Syst Appl. 2009;36(7):10604-10611.

[11] Han J, Kamber M, Pei J. Data Mining. Burlington: Elsevier Science; 2012.

[12] Gregory Mankiw N, Swagel P. The politics and economics of offshore outsourcing. J Monet Econ. 2006;53(5):1027-1056.

[13] Azam M, Chin A, Prakash N. The Returns to English-Language Skills in India. Econ Dev Cult Change. 2013;61(2):335-367.

[14] Subramanian, T. S. R. (2016). Report of the Committee for Evolution of the New Education Policy. New Delhi: Government of India.

[15] Moradpour, S., Long, S., 2017. K-mean clustering method in transportation problems, a work zone simulator case study. Proceedings of the International Annual Conference of the American Society for Engineering management. American Society for Engineering Management (ASEM).

[16] Rygielski C, Wang J, Yen D. Data mining techniques for customer relationship management. Technol Soc. 2002;24(4):483-502.

[17] Halkidi, M., Batistakis, Y. & Vazirgiannis, M. On Clustering Validation Techniques. Journal of Intelligent Information Systems 17, 107–145 (2001)

[18] Chu H, Liau C, Lin C, Su B. Integration of fuzzy cluster analysis and kernel density estimation for tracking typhoon trajectories in the Taiwan region. Expert Syst Appl. 2012;39(10):9451-9457.

⋆⋆⋆⋆⋆⋆⋆⋆⋆
ISSN(P):2319 − 3786
Malaya Journal of Matematik
ISSN(O):2321 − 5666
⋆⋆⋆⋆⋆⋆⋆⋆⋆