Self-Organizing Map vs Initial Centroid Selection Optimization to Enhance K-Means with Genetic Algorithm to Cluster Transcribed Broadcast News Documents

Ahmed Maghawry[1], Yasser Omar[1], and Amr Badr[2]
[1]Department of Computer Science, Arab Academy for Science and Technology, Egypt
[2]Department of Computer Science, Cairo University, Egypt

Abstract: *A compilation of artificial intelligence techniques are employed in this research to enhance the process of clustering transcribed text documents obtained from audio sources. Many clustering techniques suffer from drawbacks that may cause the algorithm to tend to sub optimal solutions, handling these drawbacks is essential to get better clustering results and avoid sub optimal solutions. The main target of our research is to enhance automatic topic clustering of transcribed speech documents, and examine the difference between implementing the K-means algorithm using our Initial Centroid Selection Optimization (ICSO)* [16] *with genetic algorithm optimization with Chi-square similarity measure to cluster a data set then use a self-organizing map to enhance the clustering process of the same data set, both techniques will be compared in terms of accuracy. The evaluation showed that using K-means with ICSO and genetic algorithm achieved the highest average accuracy.*

Keywords: *Clustering, k-means, self-organizing maps, genetic algorithm, speech transcripts, centroid selection*.

Received May 21, 2017; accepted July 10, 2018
https://doi.org/10.34028/iajit/17/3/5

1. Introduction

1.1. A World of Massively Growing Audible News

The larger part of news segments broadcasted on television, radio stations and on the internet are all sharing the same audible feature and are all growing rapidly, such a rapid growth produces massive amounts of data that must be organized and stored properly in order to facilitate future search and retrieval, reliable and robust techniques are needed to organize and store these massive amounts of data. Many challenges confront the field of multimedia information retrieval despite its rapid advance in the past decade. One of the main problems challenging researchers on this field is the asymmetric nature of audio and video. As regards audio, two main directions where focused on during the analysis of audio documents. The first direction was to develop audio data classification schemes to segment an audio document into coherent chunks of different types of audio classes- music, speech, speech and music etc., [13, 20, 23].

2. Background

2.1. K-Means Clustering Algorithm

The K-means clustering algorithm will be used in this research, not only because it's one of the most commonly used clustering techniques but also because it has been applied in many scientific and technological fields [6, 19, 27]. The K-means method has not only suffered from a major problem of which the algorithm produces empty clusters [3] added to that the problem produced by the random nature of cluster's initial centres selection that causes the algorithm to tend to sub optimal solutions [17]. K-means clustering algorithm will be used to group transcribed textual documents obtained from audio sources into topics by applying a similarity measure based on the Chi-square method, which is designed to eliminate non informative words that will more likely be erroneous words when applied on transcribed documents [5]. The K-means clustering algorithm belongs to the partitioning based and non-hierarchical clustering techniques [1]. The algorithm starts with a set of numeric objects X and an integer number k, then attempts to find the partition of X into k clusters while minimizing the sum of squared errors [8]. First the K-means algorithm initializes the k cluster centres. Second, the algorithm attempts to allocate each of the input data points to the closest centres according to the square of the Euclidean distance from the cluster [21]. Third, the mean value of each cluster is computed in order to update the cluster centre. This updating process happens because of the change in the membership of each cluster [26, 28]. Re-assigning the membership of the input vectors and the continuous

update of the cluster centres is repeated until no more changes in the value of any of the cluster centres occurs. K-means is commonly used because of its simplicity and the ability of applying it on a wide variety of data types. However, it's quite sensitive to the initial positions of cluster centres. Listed below are the steps of the K-means algorithm 1. Initialization: K data points are chosen randomly to initialize the K cluster centres 2. Nearest-neighbour search: for each data point, the data point will be assigned to a cluster centre if this cluster centre is the closest to that data point. How near the data vector is close to a centroid is calculated using Equation (1).

$$d(z_p , a_j) = \sqrt{\sum_{k=1}^{d} (z_{pk} - a_{jk})^2} \quad (1)$$

Where *d* represents the dimension of the data point vector, *Zp* represents the centroid of the cluster *P* and *aj* is the data point's vector.3. Updating the mean: for each cluster, calculate the mean of the input vectors assigned to that cluster to find the new cluster's centre.4. Stopping criteria: step 2 and step 3 are repeated until there's no change in the value of the calculated means.

2.2. Genetic Algorithm

On the other hand, genetic algorithms where introduced by Holland [2] and further described by Goldberg [7] as optimization technique to search for global or near global optimal solutions, it's a smart exploitation of the random search used to solve optimization problems. To overcome the transcription errors produced by the common drawbacks of Automatic Speech Recognition (ASR), root-based stemming technique is applied. To achieve topic identification, K-means [24] clustering technique is utilized.

2.3. K-means with GA and Optimized Initial Centroid Selection

This work embraces the approach of applying ASR technology to Arabic news audio documents, and then applies pre-processing techniques [9] and clustering algorithm on the transcribed textual documents produced by the ASR as in Figure 2, and then attempt to optimize the operation of the K-means initial centroid selection using Initial Centroid Selection Optimization (ICSO) an approach presented in this research, which should enhance the quality of the randomly selected centroids as in Figure 4. Finally introduce these centroids for the K-means algorithm and produce a number of clustering solutions, and deliver these solutions as the initial population for the genetic algorithm to attempt to find the global or near global optimal solution [21, 25].

2.4. Self-Organizing Maps

Given a grouped vectors in an input space, a Self-Organizing Map (SOM) will learn how to classify new input vectors [15]. Unlike competitive layers in neighbouring neurons, a self-organizing map will learn how to recognize neighbouring sections of the input space [9]. Just as competitive layers, self-organizing maps will learn the distribution; moreover it will also learn the topology of the input vectors they are trained on. SOM is focused on in this research to perform document clustering, SOM is preferred over other clustering techniques for a couple of reasons, first it preserves topology, second, clustering is performed nonlinearly on the given input data set. The topology preserving feature presented by SOM allows it not only to group similar documents together in a cluster but also organize similar clusters close together and that's unlike many other clustering methods. We constructed a 1-D SOM neural network that will receive the generated document vector as input [14].

The size of the network in terms of the number of hidden neurons is based on the desired number of clusters. The network is trained using the input document vector for about 200 epochs. The network will output the weights the centres of each cluster. Then we assign each document to its appropriate cluster for evaluation.

2.5. K-means with GA and ICSO VS SOM

The ICSO step will be applied on a subset of the data set to get randomly selected optimized initial centroids, this subset of the data set will be already clustered so the chosen initial centroids will be classified once selected to indicate to which cluster each belong to. This data will be combined together and passed to the Self-Organizing Map to be trained, then suspend the ICSO and use the trained Self-Organizing Map to generate the initial random centroids. The SOM will be used to determine a vector's membership in a certain class given its weights. Since the training data is labelled this will be a supervised learning. The process of training the SOM will be as in Figure 1.

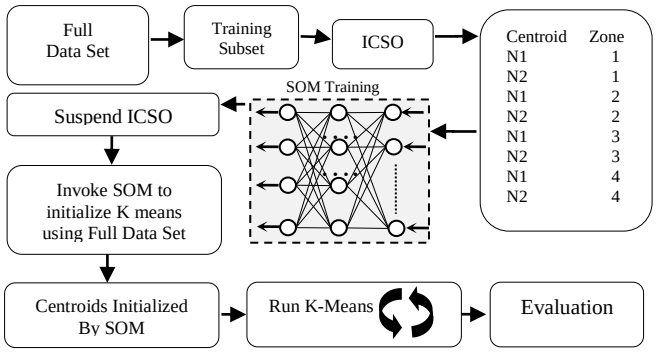

Figure 1. Neural network training process.

Our results showed the effectiveness of ICSO in initializing initial centroids for K-means, thus ICSO will be used to train the Self-Organizing Map. The topic clustering accuracy is evaluated for the selected

clustering algorithm in six situations: When the transcribed documents are clustered using pure K-means without the use of ICSO or SOM or Ga, when clustered with ICSO support, and when clustered using ICSO and Genetic Algorithms (GA) optimization, when clustered using K-means with SOM, and finally K-means with SOM and GA as shown in Table 1.

Table 1. Testing scenarios.

Case ID	K-MEANS	ICSO	SOM	GA
A	✓	×	×	×
B	✓	✓	×	×
C	✓	×	×	✓
D	✓	✓	×	✓
E	✓	×	✓	×
F	✓	×	✓	✓

2.6. SOM Description

The proposed SOM will consist of an input layer followed by four hidden layers one for each class, and the output layer, the input layer will consist of N neurons each neuron will represent a weight of a word in the vector, each hidden layer will process the inputs and will deliver a value to the output layer that consists of four neurons which will output the degree of membership of a vector in each of the four classes.

All techniques mentioned above will be a part of the main proposed model that aims to perform a clustering-based topic identification of transcribed textual files obtained out of audio files as in Figure 2.

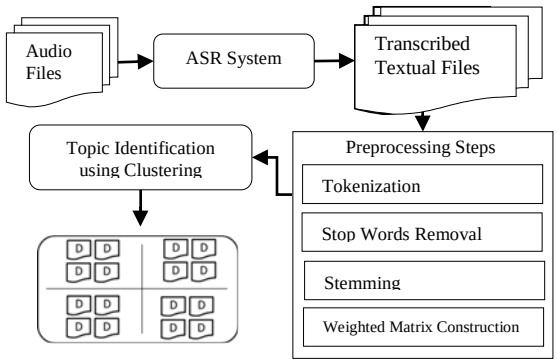

Figure 2. Overall clustering process.

3. K-Means Challenges

Despite the simplicity of k-means and its wide scale of usage in different fields, there are some challenges related to it, one of the most important drawbacks of k-means is that the algorithm's final clustering result is extremely sensitive to one of the basic and mandatory steps of k-means which is the initial random centroids selection [12]. As a result, for a given clustering problem, different algorithm runs can output different clustering solutions for the same problem depending on the initial centroids selected [6], that's why in many previous researches and applications, k-means results in terms of accuracy where not on the top because the algorithm may tend to sub optimal solutions [10]. Therefore we propose in this paper that if we provided the k-means algorithm with high quality initial centroids, the algorithm will show significant results. Furthermore, if genetic algorithm optimization was applied alongside k-means with high quality initial centroids, the algorithm will show results that might exceed other clustering techniques.

4. Proposed Model

Final results of K-means is affected by the initial selection of random centroids, the algorithm depends on the initial centroids to compute distances between them and the data set objects targeted for clustering and assign each object to its closest centroid, then calculate the mean of each formed cluster. Suppose we have a sample data set of 100 elements that we already know that they can be divided into 4 clusters each containing 25 elements. Passing this data set to the k-means algorithm to cluster it and initialize the algorithm with k=4, Four random centroids will be picked as the algorithm starts, the problem appears when the algorithm picks more than one centroid that should belong to the "same" class, furthermore, the algorithm might pick all 4 initial random centroids from the same category, because of that, obviously the algorithm will out put a solution with very bad accuracy. That's why the Initial Centroid Selection Optimization technique presented in this paper will be used to guide the k-means algorithm to pick initial centroids suitable for the K-means to start with, and later suspend ICSO and use Self-Organizing Map for the same objective, both on ICSO and SOM should achieve our goal in this step, in other words, maximize the probability that the algorithm will pick 4 initial random centroids that doesn't belong to the same cluster. Our data set will be text transcripts gained from transcribing Arabic audio news files.

4.1. Vector Representation Model

Initially, all files will be represented using Vector Representation Model (VRM) [10]. Then, these vectors will be sorted by each word's weight either ascending or descending both will be the same and both will achieve our objective which is, by sorting them, those vectors who are similar will be grouped together. Each vector is a row matrix 1xn where n is the number of all unique words that are present in all the transcribed files, and all words will be grouped into zones within the vector so the summation of all the weights that belong to a specific zone will describe the weight of each document regarding that zone as in Figure 3.

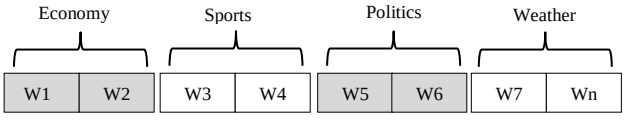

Figure 3. Vector composition.

4.2. Initial Centroid Selection Optimization

All vectors will be divided into k groups derived from the user specified k, and the K-means algorithm will be directed to choose the k initial random centroids one from each group. As a result, we will maximize the probability that each initial centroid will be more likely different than the others hence doesn't belong to the same cluster, hence provide high quality initial centroids to the k-means algorithm to start with as visualized in Figure 4.

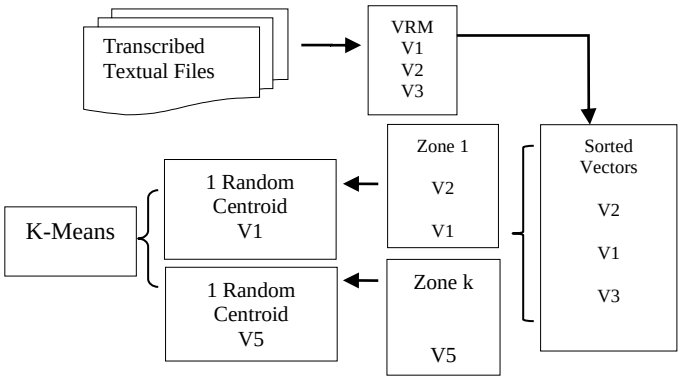

Figure 4. Providing quality centroids.

After the *n* vectors are sorted, they will be divided using Equation (2):

$$Z = {}^{n}/_{k} \quad (2)$$

Where n is the total number of vectors and k is the user specified number of clusters and *Z* is the number of zones.

4.3. Genetic Algorithm to Optimize K-Means

A genetic algorithm is a randomized search and optimization technique which is guided by principles of evolution and natural genetics, having a large amount of parallelism [4, 22]. For genetic algorithm based data clustering to be applied we first need to indicate how an individual (possible solution) will be represented then initialize starting population then deliver it to an evaluation (fitness) function then select fit chromosomes then apply crossover to combine good solutions together in search of a better solution. Finally apply mutation to avoid trapping chromosomes into a local minimum value in one of its genes. Each individual represents one feature subspace. An individual's fitness represents the clustering result indicating how good it is regarding the feature space that the individual represents. The larger the fitness, denser the data in such feature subspace, the better the clustering results will be [11].

Algorithm 1: Final clustering algorithm

- *Input:*
  - *P: Population size.*
  - *PM: Population means.*
  - *K: Number of clusters.*
  - *D: Data set in VRM.*
  - *MaxGen: Maximum number of generations.*
  - *TSSD: Targeted Sum of squared distances.*
  - *TAA: Targeted Average Accuracy*
- *Output:*
  - *Result: The fittest chromosome.*
  - *Mean: mean of the fittest chromosome*
- *Steps:*

START:
#Sort the vectors either ascending or descending
For (i = 0 : P)
{
Generate K random optimized centroids using (ICSO)/SOM Deliver to K-Means forP clustering solutions.
For each P, Loop until convergence
Save each result and its updated means in P and PM at the same index.
}
Pass the P solutions gained from step 1 to the genetic algorithm as the initial population
Foreach(individual in P)
{
(a) Calculate the fitness of each individual with each mean in PM
- Result = most fit individual
- Means = means of Result
(b) If (MaxGen || TSSD ||TAA)
{
Go toEND
Deliver Result and Means as the optimal solution.
}
Else:
{
(c) Apply selection
(d) Apply Crossover
(e) Apply mutation.
(f) - Pass off spring to (a) –Loop
}
}
END:

4.4. Algorithm Explanation

Provide the algorithm with the following inputs:

1. Population size.
2. Number of clusters.
3. Data set.
4. Maximum number of generations.
5. Targeted sum of square distances.
6. Targeted Average Accuracy.

The first step is to sort all the vectors either ascending or descending, by doing that we assume that vectors with close characteristics will be grouped together. The second step is to generate K initial random centroids using the initial centroid selection optimization method shown in Figure 5.

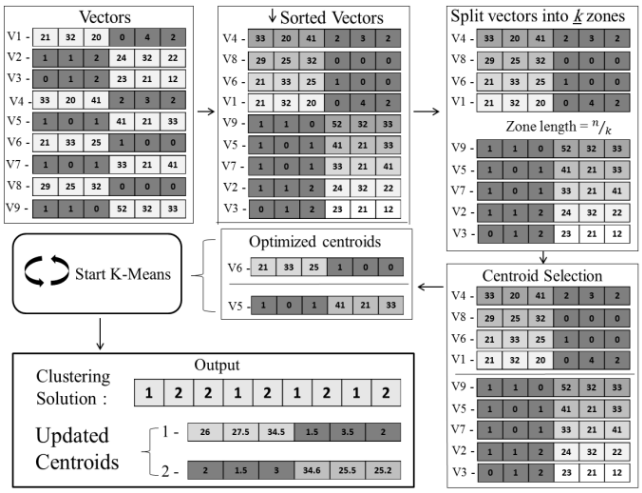

Figure 5. Quality centroid selection.

We assume that by using ICSO method to generate k initial centroids for each chromosome of P, and running k-means to produce P clustering solutions and update each cluster centre until convergence as in Figure 6, all clustering solutions obtained in this step will already be a decent clustering results that might encounter the issue of not being the optimal solution, thus, deliver them to genetic algorithm for optimization.

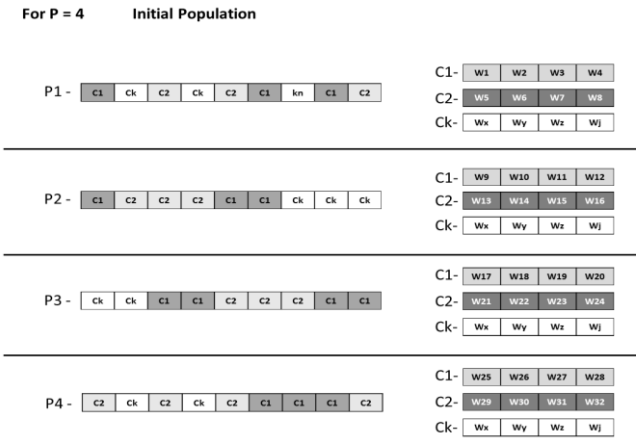

Figure 6. Initial population with centroids.

The third step, each *P* and *PM* will be concatenated for each solution at *P* to form the final structure of the chromosome (possible solution) as shown in Figure 7.

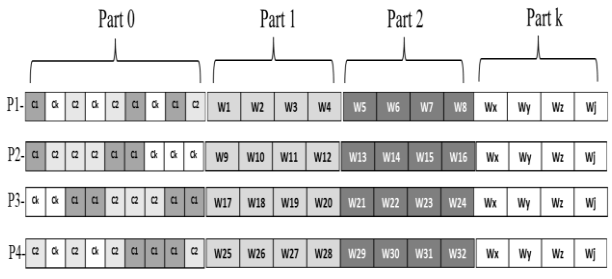

Figure 7. Final chromosome structure.

Now that we have P number of clustering solutions concatenated with their updated centroids obtained from the previous step, the third step is to deliver these chromosomes to the genetic algorithm to operate on them to attempt to search for the most optimum clustering solution.

Fourth, the genetic algorithm will compute the fitness of all chromosomes in terms of average accuracy by evaluating whether the algorithm assigned all documents to the right clusters and average SSD by calculating the sum of square distances between cluster elements and their centroid as in Equation (3):

$$f(C1, C2, \dots, Cn) = \sum_{i=1}^{K} \sum_{x_j \in C_i} ||x_j - z_i|| \quad (3)$$

Then keep the fittest chromosome and its means, and then apply genetic operators only on parts from 1 to k as in Figure 7 of each chromosome. To maintain chromosome integrity during the crossover operation we must crossover corresponding parts of each chromosome for example "part 1 from chromosome 1 with part 1 from chromosome 2", because each corresponding parts are generated from the same zone in the initial centroid selection optimization phase as shown in Figure 5. That's why we assume that part 1 from a chromosome is at the same context with part 1 from another chromosome.

Then deliver the offspring to the fitness function and loop until maximum number of generations or Targeted Sum of Squared Distances (TSSD) or Targeted Average Accuracy (TAA) is reached. Finally we will acquire a clustering solution to a problem to calculate the average accuracy and to compare it to previous results.

4.5. SOM Training

Several runs of ICSO will be performed to get multiple randomly generated centroids from 18% of the total data set to train SOM be trained and then suspend ICSO and use SOM to generate optimized initial centroids for the K-means then apply GA, finally the results acquired using SOM will be compared with that of ICSO.

5. Experimental Results Evaluation

The proposed algorithm and techniques will be tested on a data set combined of 1000 transcribed Arabic news broadcast videos, 18% of the transcripts where categorized into 4 sets of news categories (Politics, Weather, Business, Sport), then a collection of text files pre-processing procedures were made on them as following: Tokenization, word grouping, words suspension, all these steps are done on the data set to prepare it to be presented in Vector Representation Model to get the weighted matrix of all documents regarding the constructed vector as in Figure 8, then start our implementation.

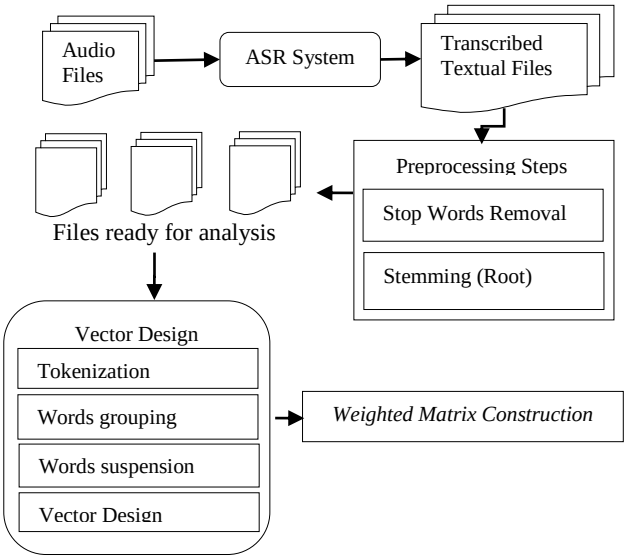

Figure 8. Dataset to VRM conversion.

Now that we got the vector designed and words from the same categories grouped together in regions within that vector as mentioned before, the remaining 82% of the data set will be represented in VRM using the same vector construction. Four different algorithms of K-means were used [18], (Lloyd, Forgy's, McQueen, Hartigan-Wong), advantages and disadvantages of each is listed in Table 2.

Table 2. K-means algorithms.

Algorithm	Advantages	Disadvantages
Lloyd	- For large data sets - Discrete data distribution - Optimize total SSD	- Slower convergence - Possible to create empty clusters
Forgy's	- For large data sets - Continuous data distribution - Optimize total SSD	- Slower convergence - Possible to create empty clusters
McQueen	-Fast initial convergence - Optimize total SSD	- Need to store the two nearest-cluster computations for each case - Sensitive to the order the algorithm is applied to the cases
Hartigan - Wong	- Fast initial convergence - Optimize within-cluster SSD	- Need to store the two nearest-cluster computations for each case - Sensitive to the order the algorithm is applied to the cases

All files of the data set where randomly shuffled and given a standard name from D1 to Dn where n is the total number of documents and the following test cases were applied. First test case, a pure K-Means was applied on the data set which we know in advance that it has Business, Politics, Sport, Weather, 250 file in each category. The Second test case was applying K-Means with centroid optimization on the same data set. The Third was applying K-Means with genetic algorithm optimization. The Fourth was applying K-Means with initial centroid selection optimization and genetic algorithm optimization, then Apply K-means with SOM. Finally apply K-means with SOM and GA, all previous scenarios are repeated 4 times, one for each version of k-means, all previous test cases where executed on an Intel CORE i7-5600 @ 2.60 GHz 2 Core 4 Thread 16 GB Main Memory 64 bit Win 7 Enterprise and the following results were acquired:

After applying K-Means only:

Table 3. Clustering using k-means only.

#	K-Means Algorithm	Avg Accuracy	Avg Iterations	Avg SSD
1	Hartigan-Wong	83.3%	2.67	10714603.925
2	Lloyd	76%	3.6	9155058.925
3	Forgy	80.67 %	4	9210684.500
4	McQueen	83.3 %	3.3	4857707.525
-	Average SSD	3,393,8054.875		
-	Average Accuracy	80.81%		
-	Average Iterations	3.39		

After applying K-Means with GA.

Table 4. Clustering using K-means with GA.

#	K-MeansAlgorithm	Avg Accuracy	Avg MaxGen	Avg Iterations	Avg SSD
1	Hartigan-Wong	84.7%	17	2.67	10524203.67
2	Lloyd	78%	24	3.6	9155058.49
3	Forgy	81.41%	28	4	9210622.500
4	McQueen	84.3%	18	3.3	4857693.525
-	Average SSD	3,374,7617.912			
-	Average Accuracy	82.10%			
-	Average MaxGen	21.75			
-	Average Iterations	3.39			

Applying K-Means with initial centroid selection optimization:

Table 5. Clustering using k-means with initial centroid selection optimization.

#	K-Means Algorithm	Avg Accuracy	Avg Iterations	Avg SSD
1	Hartigan-Wong	100%	1.67	2115090.21
2	Lloyd	86.3%	2.6	2238123.64
3	Forgy	84.67 %	2.6	2238123.64
4	McQueen	86.3%	1.67	2238123.64
-	Average SSD	2,207,365.28		
-	Average Accuracy	89.31 %		
-	Average Iterations	2.135		

Applying K-Means with initial centroid selection optimization and GA:

Table 6. Clustering Using K-means with GA and ICSO.

#	K-Means Algorithm	Avg Accuracy	Avg MaxGen	Avg Iterations	Avg SSD
1	Hartigan-Wong	100%	13	1.67	2115046.21
2	Lloyd	86.3%	21	2.6	2238075.24
3	Forgy	84.67 %	25	2.6	2238075.24
4	McQueen	86.3%	14	1.67	2238075.24
-	Average SSD	2,207,317.98			
-	Average Accuracy	89.31 %			
-	Average MaxGen	18.25			
-	Average Iterations	2.135			

K-means with SOM:

Table 7. Clustering Using K-means with SOM.

#	K-MeansAlgorithm	Avg Accuracy	Avg Iterations	Avg SSD
1	Hartigan-Wong	86%	2.14	2554329.82
2	Lloyd	82.7%	3.1	2677358.85
3	Forgy	83.67 %	3.5	2677358.85
4	McQueen	84.2%	2.9	2677358.85
-	Average SSD	2,646,601.59		
-	Average Accuracy	84.14 %		
-	Average Iterations	2.91		

K-means with SOM and GA:

Table 8. Clustering Using K-means with SOM and GA.

#	K-MeansAlgorithm	Avg Accuracy	Avg MaxGen	Avg Iterations	Avg SSD
1	Hartigan-Wong	87.2%	16	2.14	2224329.82
2	Lloyd	83.7%	23	3.1	2309358.85
3	Forgy	84.14 %	27	3.5	2309358.85
4	McQueen	85.21%	17	2.9	2309358.85
-	Average SSD	2,288,101.59			
-	Average Accuracy	84.14 %			
-	Average MaxGen	20.75			
-	Average Iterations	2.91			

5.1. Final Results

Results from Tables 3, 4, 5, 6, 7, and 8 are acquired using "speechnotes" [28] transcriber with WER=10.2% for 40435 reference words on our data set that consists of 1000 transcribed audio files using "speechnotes".

Table 9. Final results.

Test Case#	Avg SSD	Avg Acc	Avg Max Gen	AvgIter
1	33,938,054.875	80.81%	-	3.39
2	33,747,617.912	82.10%	21.75	3.39
3	2,207,365.28	88.31%	-	2.135
4	2,207,317.98	88.31%	18.25	2.135
5	2,646,601.59	84.14 %	-	2.91
6	2,288,101.72	84.14 %	20.75	2.91

Following is graph visualization for the test results regarding the average sums of squared distances for all test cases are plotted against each other in Figure 9.

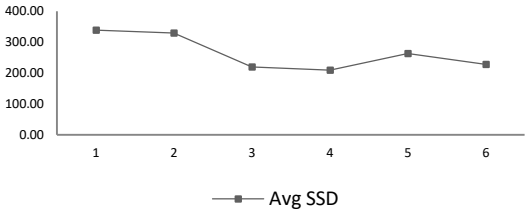

Figure 9. Average sum of squared distances for all test cases.

The average accuracy of the previous test cases is plotted against each other in Figure 10.

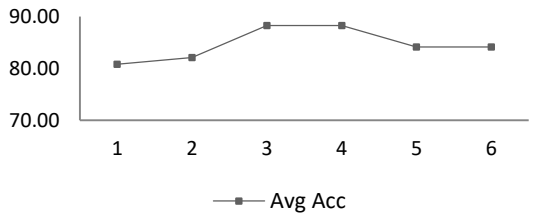

Figure 10. Average accuracy.

The average maximum number of generations for all test cases is plotted against each other in Figure 11.

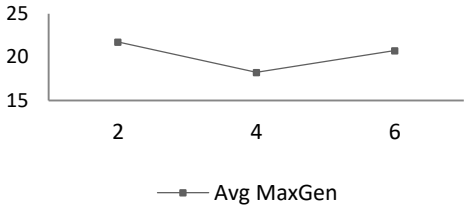

Figure 11. Average maximum number of generations.

The average number of iterations for all test cases is plotted against each other in Figure 12.

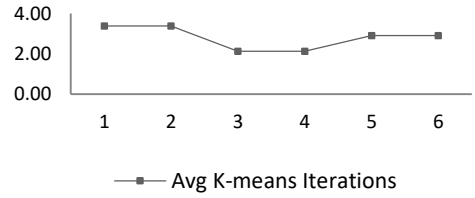

Figure 12. Average number of K-means iterations.

Previous results [10]: Previous results acquired using Dragon Dictation Recognition System with WER=20.6 for 30040 reference words on previous data set consisting of 1000 transcribed audio files using Dragon Dictation Recognition System shown in Table 10.

Table 10. Results from applying the previous technique on previous dataset.

Clustering Approach	Average Accuracy Chi-Square
K-means	79.05%
Spectral	87.21%

Results on Previous Data Set: Our technique was applied on previous data set used in [10], results shown in Table 11.

Table 11. Results from applying the proposed technique on Previous Dataset.

Clustering Approach	Average Accuracy Chi-Square
K-means + GA + ICSO	87.91%
K-means + GA + SOM	85.06 %

6. Conclusions

Six test cases were implemented and the results were acquired in Tables 3, 4, 5, 6, 7, and 8 and then compared against each other in Table 9. Comparison shows that clustering using genetic algorithm has slightly improved the average accuracy by 1.29% and the average SSD by 190,436.963 with no change in the average iterations, while the dramatic change appeared when the initial centroid selection optimization technique was applied, which improved the average accuracy by 7.5% and the SSD by 31,730,689.595 and the average iterations by 1.25 iteration. Applying genetic algorithm after the ICSO technique has slightly improved the SSD by 47.3, but neither improved the average accuracy nor the average iterations. Finally, the ICSO technique has been used to train a Self-Organizing Map then the ICSO was suspended and replaced with the SOM which when used to pick the initial random centroids for K-means has improved the average SSD by 31,291,453.285 and the average iterations by 0.48 and the average accuracy by 3.33%, not the best results in our research but obviously better than the pure K-means, applying genetic algorithm after that has resulted in the same behaviour, with a slight enhancement in terms of average SSD by 358,499.87

We conclude that the improving impact of genetic algorithm on k-means is not as dramatic as initial centroid selection optimization, applying genetic algorithm optimization with k-means alone has resulted in a slight improvement in terms of average accuracy and the sum of square distances, while applying initial centroid selection optimization alone with k-means has resulted in a significant improvement in terms of average accuracy and average iterations for k-means to converge, finally applying genetic algorithm on the results of k-means and initial centroid selection optimization has resulted in a slight improvement in terms of sum of square distances. Finally running k-means with ICSO and GA on the dataset of [10] has resulted-as expected- in a significant improvement in terms of accuracy by 8.86%, while slightly exceeded the spectral algorithm implementation by 0.7%, While the impact of applying SOM to initialize K-means was not ranked as number one in our tests in terms of our evaluation criteria, but it achieved better performance than the previous results in comparison with their K-means implementation [10], but did not exceed the average accuracy of their Spectral clustering implementation.

7. Future Work

In this research we found out that ICSO had a dramatic impact in terms of accuracy and SSD for the clustering process while Genetic algorithm didn't had as much impact as ICSO. Future research will be conducted in order to find better ways to utilize GA to get the most efficient use for the algorithm to optimize k-means results.

References

[1] Abhishekkumar K. and Sadhana C., "Survey Report on K-Means Clustering Algorithm," *International Journal of Modern Trends in Engineering and Researc*h, vol. 4, pp. 218-221, 2017.

[2] Affenzeller M., Wagner S., and Winkler S., "Aspects of Adaptation in Natural and Artificial Evolution," *in Proceedings of the 9th Annual Conference Companion on Genetic and Evolutionary*, London, pp. 2595-2602, 2007.

[3] Agarwal S., "Data Mining: Data Mining Concepts and Techniques," *in Proceedings of International Conference on Machine Intelligence and Research Advancement*, Katra, pp. 203-207, 2013.

[4] Banerjee A. and Louis S., "A Recursive Clustering Methodology Using A Genetic Algorithm," *in Proceedings of IEEE Congress on Evolutionary Computation*1, Singapore, pp. 66-71, 2007.

[5] Coden A. and Brown E., "Speech Transcript Analysis for Automatic Search," *in Proceedings of the Hawaii International Conference on System Sciences*, Maui, pp. 9, 2001.

[6] Evritt B., Landau S., and Leese M., *Cluster Analysis*, Wiley Series in Probability and Statistics, 2011.

[7] Goldberg D., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing, 1989.

[8] Hamerly G. and Drake J., *Partitional Clustering Algorithms*, Springer, 2014.

[9] Herrmann M., "Self-Organizing Feature Maps with Self-Organizing Neighborhood Widths," *in Proceedings of ICNN95-International Conference on Neural* Networks, Perth, pp. 2998-3003, 1997.

[10] Jafar A., Fakhr M., and Farouk M., "Enhanced Clustering-Based Topic Identification of Transcribed Arabic Broadcast News," *The International Arab Journal of Information Technology*, vol. 14, no. 5, pp. 721-728, 2017.

[11] Jian-Xiang W., Huai L., Yue-Hong S., and Xin-Ning S., "Application of Genetic Algorithm in Document Clustering," *in Proceedings of International Conference on Information Technology and Computer Science*, Kiev, pp. 145-148, 2009.

[12] Joshi K. and Nalwade P., "Modified K-Means for Better Initial Cluster Centers," *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 7, pp. 219-223, 2013.

[13] Li D., Sethi I., Dimitrova N., and Mcgee T., "Classification of General Audio Data for Content-Based Retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533-544, 2001.
[14] Liu Y., Liu M., and Wang X., *Applications of Self-Organizing Maps*, Magnus Johnsson, 2012.
[15] Lu S., "Pattern Classification Using Self-Organizing Feature Maps," *in Proceedings of International Joint Conference on Neural Networks*, San Diego, pp. 471-480, 1990.
[16] Maghawry A., Omar Y., and Badr A., "Initial Centroid Selection Optimization for K-means with Genetic Algorithm to Enhance Clustering of Transcribed Arabic Broadcast News Documents," *Computational Methods for Systems and Software CoMeSySo: Applied Computational Intelligence and Mathematical Methods*, Szczecin, pp. 86-101, 2017.
[17] Mai X., Cheng J., and Wang S., "Research on Semi Supervised K-Means Clustering Algorithm in Data Mining," *Cluster Computer*, vol. 22, pp. 3513-3520, 2019.
[18] Morissette L. and Chartier S., "The K-Means Clustering Technique: General Considerations and Implementation in Mathematica," *Tutorials in Quantitative Methods for Psychology*, vol. 9, no. 1, pp. 1524, 2013.
[19] Shrivastava P., Kavita P., Singh S., Shukla M., "Comparative Analysis in Between The K-Means Algorithm, K-Means Using with Gaussian Mixture Model and Fuzzy C Means Algorithm," *in Proceedings of the International Conference on Communication and Computing Systems*, Taylor and Francis Group, London, pp. 1037-1042, 2016.
[20] Speechnotes. [Online]. Available: https://speechnotes.co/, Last Visited, 2017.
[21] Sun H. and Xiong L., "Genetic Algorithm-Based High-dimensional Data Clustering Technique," *in Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery*, Tianjin, pp. 485-489, 2009.
[22] Tiwari A., Sharma L., and Krishna G., "Entropy Weighting Genetic K-Means Algorithm for Subspace Clustering," *International Journal of Computer Applications*, vol. 7, no. 7, pp. 27-30 2010.
[23] Wold E., Blum T., Keislar D., and Wheaten J., "Content-based Classification, Search, and Retrieval of Audio," *IEEE Multimedia*, vol. 3, no. 3, p. 27-36, 1996.
[24] Wong C., "A Short Survey on Data Clustering Algorithms," *in Proceedings of 2nd International Conference on Soft Computing and Machine Intelligence*, Hong Kong, pp. 64-68, 2015.
[25] Wu J., *Advances in K-Means Clustering*, Springer, 2012.
[26] Xiao-Feng L., Kun-Qing X., Fan L., and Zheng-Yi X., "An Efficient Clustering Algorithm Based on Local Optimality of K-Means," *Journal of Software*, vol. 19, no. 7, 2008.
[27] Xu R. and Wunschii D., "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.
[28] Yadav A. and Singh S., "An Improved K-Means Clustering Algorithm," *International Journal of Computing Academic Research*, pp. 88-103, vol. 5, no. 2, 2016.

Ahmed Maghawry software developer at a leading company in Egypt in the field of electronic payments and solutions. His research interests are in artificial intelligence machine learning, and computing algorithms, received MSc in computer science from the Arab Academy for Science and Technology and Maritime Transportation.

Yasser Omar assistant professor in the Department of Computer Science, Faculty of Computing and Information Technology, Arab Academy for Science Technology & Maritime Transport. His research interests are bioinformatics, medical imaging, data visualization, machine learning, and computing algorithms. Omar received a PhD in biomedical engineering from Cairo University.

Amr Badr is a Professor in the Department of Computer Science, Faculty of Computers and Information, Cairo University. He received his BSc in Engineering with Honors in 1986. He received his MSc and PhD in Computer Science from Cairo University in 1995 and 1998. His research interests are Intelligent Systems, Bioinformatics, Medical Imaging and P-systems. He has published more than 170 journal research papers in these areas.