iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4446

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Improved Expectation Maximization (EM) Algorithm based on Initial Parameter Selection

Maria Lolita G. Masangcap[1], Ariel M. Sison[2], Ruji P. Medina[3]
[1, 2, 3] *Graduate Programs, Technological Institute of the Philippines Quezon City*

Abstract*: This paper focuses on the improved initial parameter selection for Expectation Maximization (EM) Algorithm. The initial parameters in the Standard EM Algorithm (StEM) are chosen randomly thus taking maximum time in forming clusters and leading to slow convergence. The improved initial parameter selection technique adopted the concept of firefly movement and light intensity of Firefly algorithm to better find initial parameters. A clustering comparison module was developed and different simulations using two synthetic datasets and a real dataset were conducted. The Enhanced EM Algorithm (EnEM) provided well separated initial parameters which guarantee an efficient way of assigning data points to correct clusters. All the simulations show that EnEM is constantly taking less computing time in clustering the given dataset with acceptable clustering fitness and less clustering error than StEM.

***Keywords*: Data Mining, Cluster Analysis, Expectation Maximization Algorithm, Initial Parameter Selection, Firefly Algorithm**

I. INTRODUCTION

EM algorithm is a well-established model-based clustering technique that is simple and straightforward to implement that tries to optimize the fit between the given data and some mathematical model [1] [2]. The choice of initial parameters plays an important role on the performance of the EM algorithm [3] cause it takes maximum time in forming clusters [4] that lead to slow convergence [5] and high computational cost. Therefore, these remain the need to enhance EM algorithm in its initialization stage to produce good clustering results and to take less computational time.

Since good initialization leads to fast convergence, there is a challenge of improving the initialization stage of the EM Algorithm. With that, this study aims to improve the Expectation-Maximization (EM) algorithm based on the initial parameters selection. An improved initial parameters selection technique for EM algorithm is introduced in this study which uses the concept of firefly movement and light intensity of firefly algorithm. A comparison of the clustering performance of the Enhanced EM algorithm (EnEM) to the Standard EM algorithm (StEM) in terms of clustering fitness, clustering error and computing time of the two algorithms is discussed. The application of an enhanced EM algorithm as a clustering method will open more opportunities to discover new knowledge as an outcome of a more precise data analysis.

II. RELATED LITERATURE

A. Expectation Maximization Algorithm

EM has many different applications including cluster analysis, censored data modeling, mixed models and factor analysis [6]. It is a method of cluster analysis which aims to group n observations into k clusters through the computation of maximum likelihood estimates of parameters in statistical models like Gaussian Mixture Model (GMM) [2] [4]. First, the algorithm initializes mixture model parameters by selecting data points randomly from the given data set as the initial cluster means, μ, and then compute for the variance, δ^2, and weight, w, for the dataset. Then the standard EM for GMM will perform Expectation (E) step and Maximization (M) step alternately until reaching convergence. In the E- Step, the algorithm will compute for the likelihood that each point coming from a certain cluster then estimate the probability of each cluster given a data. The data point will be assigned to the cluster with the highest probability. In the M-Step, on the other hand, an update on the parameters estimation will take place. The algorithm repeats between these two steps until reaching stopping criteria. EM has various advantages and real-world applications including banking, medical, image, etc.

Different studies were conducted in relation to EM algorithm and some includes the Expectation Conditional Maximization (ECM) which substitutes the maximization step over one's parameters of interest by conditioning on a subset of these parameters [7], the Space Alternating Generalized Expectation (SAGE) by [8] which updates the parameters in order by interchanging between several small hidden-data spaces defined by the algorithm designer. Another is the Lazy EM (LEM) that speed up the EM algorithm on the basis of partial e-steps and guaranteed to converge to strength a local maximum [9]. And the Expectation Conditional Maximization

Either (ECME) by [10] which is obtained by maximizing the likelihood function over on strength expanded parameter and tends to be a simple and effective method to accelerate its parent EM algorithm.

A good selection of initial parameters [11] or finding a good initialization [12] is a problem of EM. In [3], it is mentioned that the performance of EM relies on the selection of initial parameters. Numerous methods were already proposed in finding better initial parameters. Blömer & Bujna [13] presented new initialization methods based on the well-known K-means++ algorithm and Gonzalez algorithm. Their proposed method is designed for Gaussian mixture models which closed the gap between simple but constant initialization techniques and complex methods. And the result shows that algorithm based on K-Means++ produced good results than the other methods when compared to the initialization and between the two proposed method. Another approach was also presented by [3] which is a hybrid-based approach that aims to improve the stability of EM algorithm based on finding the optimal number of clusters and their parameters. Based on the outputs, they indicate that the approach overcomes the dependence on the initial parameters presented by the classical EM-GMM algorithm.

An enhancement of the EM algorithm is the main focus of this paper. Since the performance of EM Algorithm is strongly dependent on the choice of the initial parameters which are selected randomly [3], this study intends to improve the performance of EM Algorithm without sacrificing its stability through the elimination of the random selection of parameters in its initialization stage.

B. Firefly Algorithm

In almost all areas of optimization, firefly algorithm becomes an increasingly important tool of swarm intelligence because it is simple yet quite efficient nature-inspired search technique for global optimization that hooks the interest of researchers and developers. FA needs to be modified and hybridized in order to solve various real-world problems [14]. It is a warm-based intelligence algorithm which mimics the flashing behaviour of fireflies. A firefly attracts other fireflies when it flashes a signal that can be used for some purposes like predation or mating [15].

The firefly algorithm has two steps which includes the variation of light intensity and the calculation of attractiveness [16]. The value of light intensity changes with the distance (r) monotonically and exponentially and can be computed using equation 1. The distance is used to describe how the two fireflies are close to each other [17]. In the given equation below, I0 is defined as the initial light intensity and is defined as the light absorption coefficient.

$$I = I_0 e^{-\gamma r} \qquad (1)$$

On the other hand, equation 2 is used to compute the brightness of each firefly which will define the value of the firefly's attractiveness [18] [19]. In the equation, the distance between the two fireflies is denoted by 'r' and β_0 defines their attractiveness at $r=0$.

$$\beta = \beta_0 e^{-\gamma r^2} \qquad (2)$$

A variant of Firefly algorithm was proposed by [20]. In their modified version of FA, the weakness of the standard FA is reduced through the enhancement of the collective movement of fireflies. In their process of enhancing FA, they used the value of the global optimum in the movement of the fireflies. It is the same way how Hassanzadeh & Meybodi [21] used the concept of global optimum as the firefly which will influence others. When firefly1 is compared with firefly2 and it happens that firefly2 is much brighter then there would be a tendency for firefly1 to move towards firefly2. Firefly2 will then be considered as the global optima and will be updated in any iteration. Cartesian distance is used to compute the distance of fireflies to global optima which is determined using equation 3:

$$\beta r_{i,best} = \sqrt{(x_i - x_{gbest})^2 + (y_i - y_{gbest})^2} \qquad (3)$$

The movement of the firefly can be determined by using equation 4:

$$x_i = x_i + \left(\beta_0 e^{-\gamma r_{ij}^2}(x_j - x_i) + \beta_0 e^{-\gamma r_{igbest}^2}(x_{gbest} - x_i)\right) + \propto \left(rand - \frac{1}{2}\right) \qquad (41)$$

In equation 4, α is a randomization parameter and rand is the random number generator in which its numbers are uniformly distributed in the interval [0, 1]. β_0 is the attractiveness at $r = 0$ and γ is the light absorption coefficient at the source. The parameter

γ characterizes the variation of the attractiveness and its value is important to determine the speed of the convergence. For most cases of execution, β_0 takes the value of 1 and the value of γ varies from 0.01 to 100.

III. THE EM ALGORITHM AND ITS ENHANCEMENT

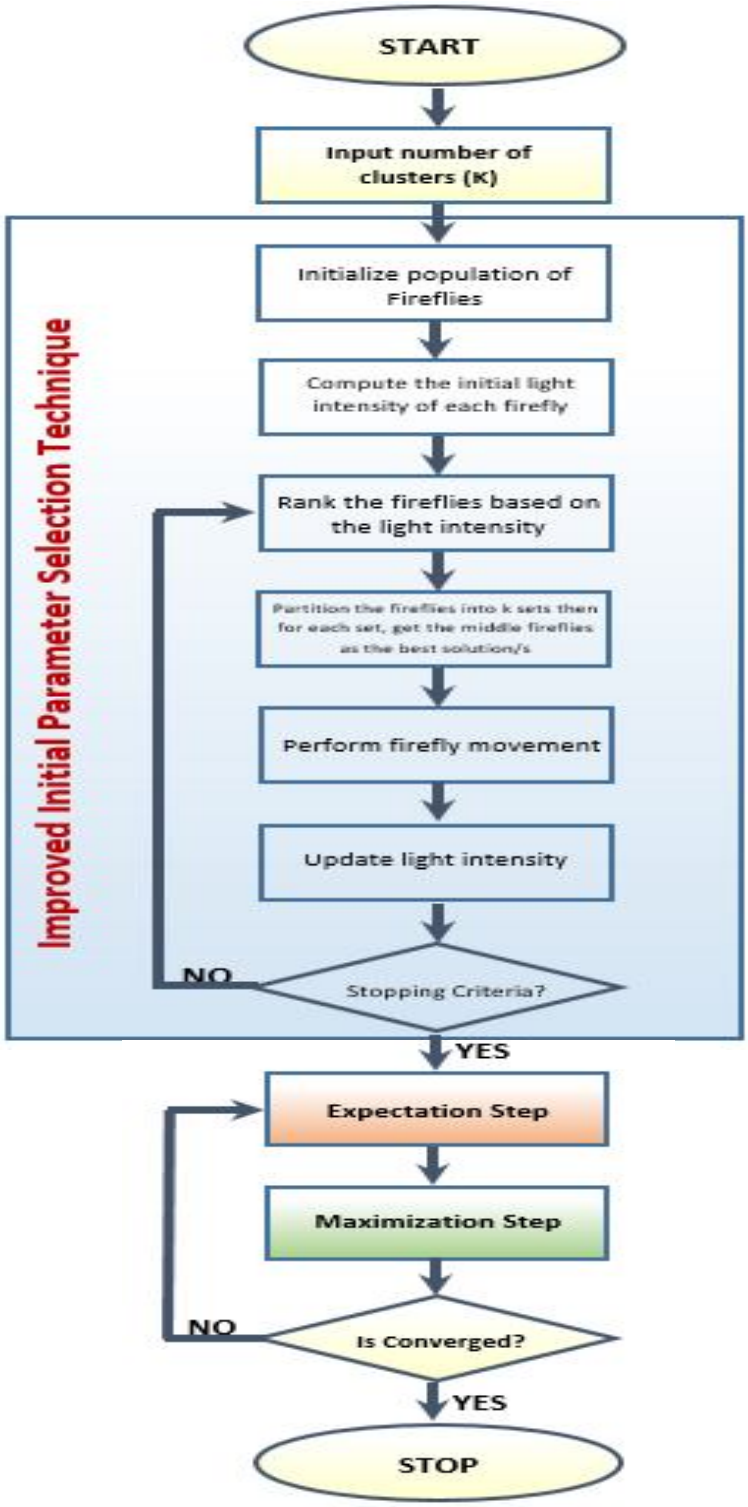


Fig. 1 Enhanced EM Algorithm

In this study, the concepts of firefly movement and light intensity of Firefly algorithm were adopted in improving the initial parameters selection of the standard EM algorithm. Figure 1 shows an improved initial parameter selection technique used in enhancing Expectation Maximization Algorithm. In the figure, an initialization of the firefly population is needed as its initial step. From the firefly population, given the data points, an initial light intensity will be computed based on the distance of each firefly to the origin. As defined, the light intensity of a firefly varies with the distance having a fixed light absorption coefficient. The attractiveness of the firefly is directly proportional to the light intensity seen by the fireflies adjacent to it [21]. During the processing, less bright firefly move towards the brighter firefly. Ranking of fireflies will be based on the light intensity.

With the given initial ranking of fireflies, the population will be partitioned into *k* number of sets. Each set can contain *n/k* data points. In each set, middle firefly will be identified as the best solutions. Given the best solutions, the distance of each firefly to the identified best solutions will be determined using the Cartesian distance using equation 3. Minimum distance will be considered in the computation of the firefly movement. This means that the movement of the firefly will use the data of the best solution near to the data point. In this proposed algorithm, movement of fireflies will also be influenced by its neighbor's distance as well as the identified best solution. The firefly's movement will be computed using equation 4. After the movement, an update to the light intensity will also be done using the new position of each firefly. Then, the ranking will again be done using the updated light intensity. This process will continue until meeting the stopping criteria. In the final ranking, the middle fireflies will be identified as the best solutions and will be selected as the initial parameters for the implementation of the EM algorithm. This process will guarantee well-separated parameters to be used in the implementation of E and M steps of the original EM Algorithm.

IV. EXPERIMENTAL EVALUATION

To determine the clustering performance of both standard EM (StEM) and enhanced EM (EnEM), comparisons were made in terms of clustering fitness, clustering error and computing time. The clustering fitness of the two algorithms was measured through the calculation of intra-cluster similarity as well as the inter-cluster similarity [22]. A good clustering algorithm aims to generate clusters with lower inter-cluster similarity and higher intra-cluster similarity resulting in a high clustering fitness. As for the measurement of the clustering error, Sum of Squared Errors (SSE) was computed for all the clustering results. A lesser value of SSE means a better or good clustering performance [2]. The aim of a good clustering technique is to lessen the within-cluster sum of square errors. The lesser the SSE, the better the goodness of the clustering algorithm. On the other hand, the clustering performance in terms of computing time will be measured in milliseconds.

An experimental evaluation was performed in order to measure the performance of the two algorithms. An EM Clustering Comparison Module was developed to simulate the two algorithms and show clustering results. The module shown in Figure 2 was designed and developed using Microsoft Visual Studio Express 2012.

The module is composed of functionalities intended to compare the standard EM algorithm and Enhanced EM Algorithm. The module's functionalities are as follow:

A. Datapoints Loader

This allows the user to input the path name of the dataset to be clustered. Number of clusters and maximum iterations are also needed inputs in the module. The button Run must be clicked in order to output the clustering results of the given dataset.

B. Clustering Results

Upon clicking the Run button, the clustering results of both standard and enhanced EM will be displayed. A scatter plot of the clustering result is also available in the module.

C. Result of Comparison

This is where the comparison of the clustering performance of the two algorithms will be shown. The result of clustering fitness, clustering error and the computing time will be presented. A line graph of the results is also displayed to make it easier to compare the clustering performance of the two algorithms.

The implementation of the two algorithms was carried out on a system with Intel Core i5 with 240 GHz processor speed, six (6) GB RAM and with a hard disk capacity of 512 Gb. Three datasets were used in the Experiment: two synthetic datasets (3000 and 5000 datapoints) generated by a rand() function in MS Excel ranging from 50.0000 to 99.9999 and one real dataset (1000 datapoints) which consists of the result of students' performance of Bataan Peninsula State University in two semesters. A clustering of five (5), seven (7), and ten (10) clusters is done on each dataset and used as input in the comparison of the clustering performance. Both

algorithms used only a univariate Gaussian distribution.

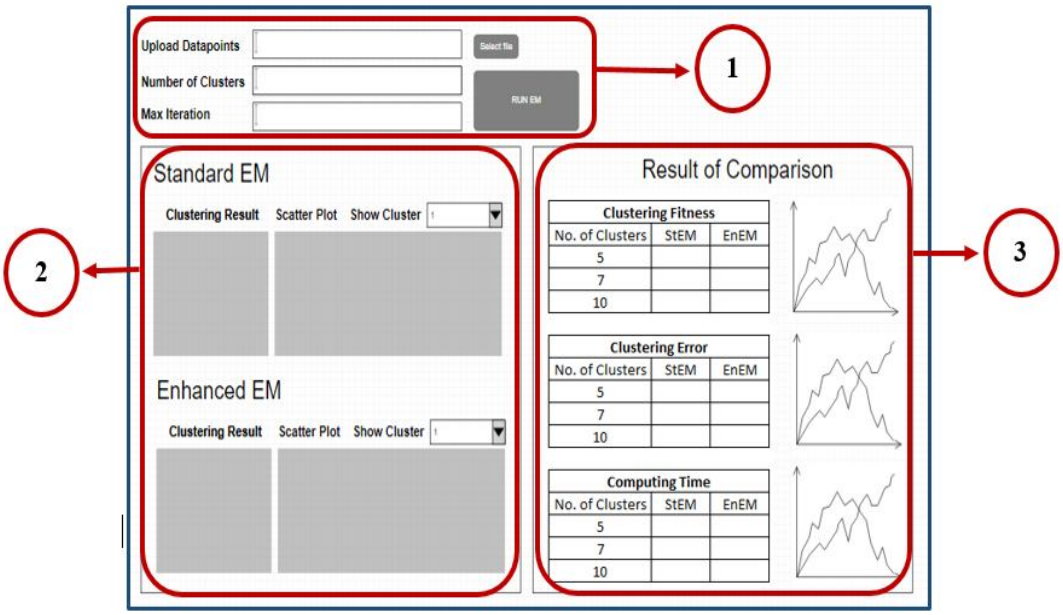


Fig. 2 EM Clustering Comparison Module Interface

V. RESULTS AND DISCUSSION

The simulation was performed through the module developed. Both algorithms were implemented using three (3) different datasets; the difference is that the Enhanced EM (EnEM) systemically chose the initial parameters using the concept of the firefly algorithm whereas the Standard EM (StEM) randomly picked the initials parameters. To better observe the clustering performance of both algorithms, results from five (5) executions were recorded and the average of these values were used.

Figure 3 and Table I show the performance of both algorithms in terms of clustering fitness. As shown, higher clustering fitness results were generated through the implementation of the Enhanced EM algorithm.

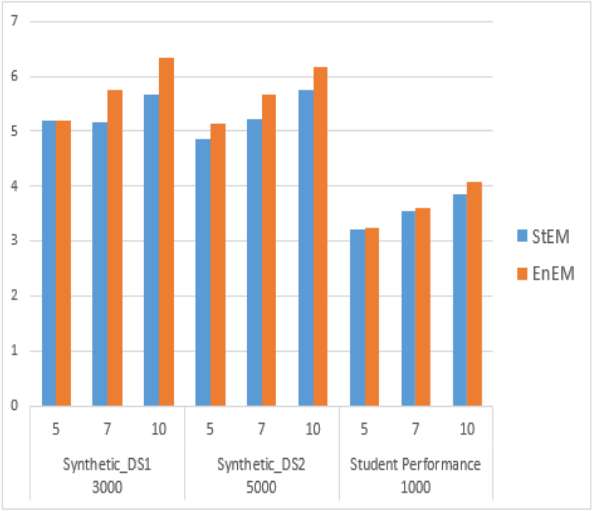


Fig. 3 Graphical Representation of Clustering Fitness

TABLE I

RESULTS IN TERMS OF CLUSTERING FITNESS

Number of Clusters	Synthetic_DS1 3000		Synthetic_DS2 5000		Student Performance 1000	
	StEM	EnEM	StEM	EnEM	StEM	EnEM
5	5.2	5.2	4.87	5.14	3.22	3.24
7	5.16	5.74	5.22	5.66	3.56	3.61
10	5.68	6.34	5.74	6.16	3.86	4.09

The comparison of clustering error results of both algorithm is shown in Figure 4 and Table II. Significant lower values for the Sum of Squared Errors were computed using the clustering results of the Enhanced EM algorithm. This means a less within-cluster sum

of squared error for the EnEM which implies a better clustering performance of EnEM as compared to StEM.

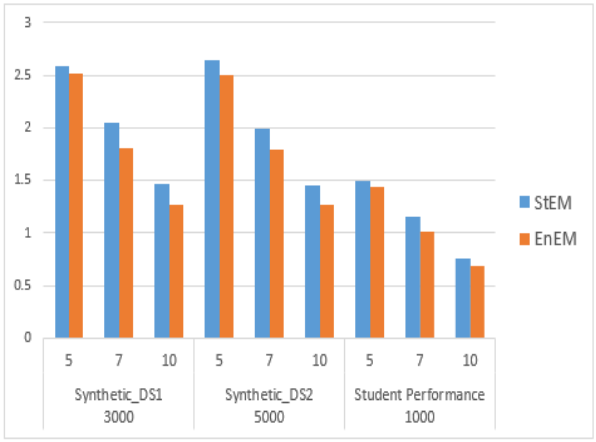


Fig. 4 Graphical Representation of Clustering Error

TABLE II

RESULTS IN TERMS OF CLUSTERING ERROR

Number of Clusters	Synthetic_DS1 3000		Synthetic_DS2 5000		Student Performance 1000	
	StEM	EnEM	StEM	EnEM	StEM	EnEM
5	2.59	2.52	2.65	2.5	1.5	1.44
7	2.05	1.81	1.99	1.79	1.16	1.01
10	1.47	1.27	1.45	1.27	0.75	0.69

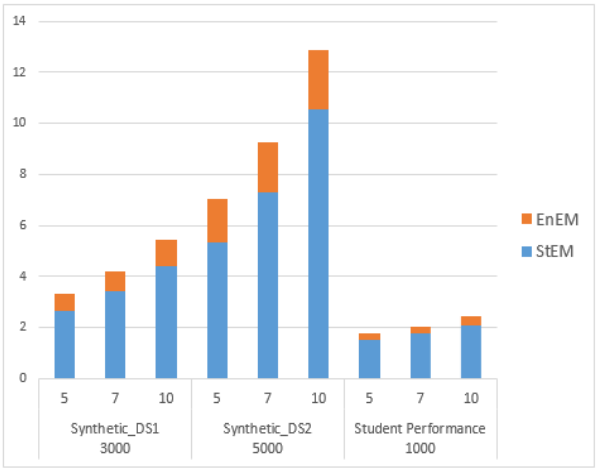


Fig. 5 Graphical Representation of Computing Time

TABLE III

RESULTS IN TERMS OF COMPUTING TIME

Number of Clusters	Synthetic_DS1 3000		Synthetic_DS2 5000		Student Performance 1000	
	StEM	EnEM	StEM	EnEM	StEM	EnEM
5	2.62	0.7	5.32	1.73	1.52	0.24
7	3.41	0.81	7.28	1.98	1.77	0.27
10	4.41	1.01	10.56	2.32	2.09	0.33

The values shown in Figure 5 and Table III present a lower computing time for the EnEM. This means the EnEM spent less time in generating clusters as presented in the simulation using three different datasets. Since a systematic way of choosing the initial parameters is implemented in EnEM, a faster EM is presented in this study.

VI. CONCLUSIONS

The improved initial parameters selection technique for EM Algorithm applying the concepts of firefly movement and light intensity of Firefly Algorithm remove the random selection of the initial parameters and reduced the number of rounds of implementing the E and M steps. The technique implemented a more systematic way of choosing initial parameters thus, guarantees well-separated parameters. An EnEM Clustering Comparison Module was developed to simulate and evaluate the clustering performance of the two algorithms. The output of the developed module shows that EnEM surpassed the performance of StEM in clustering data points in a dataset in terms of clustering fitness, clustering error and computing time.

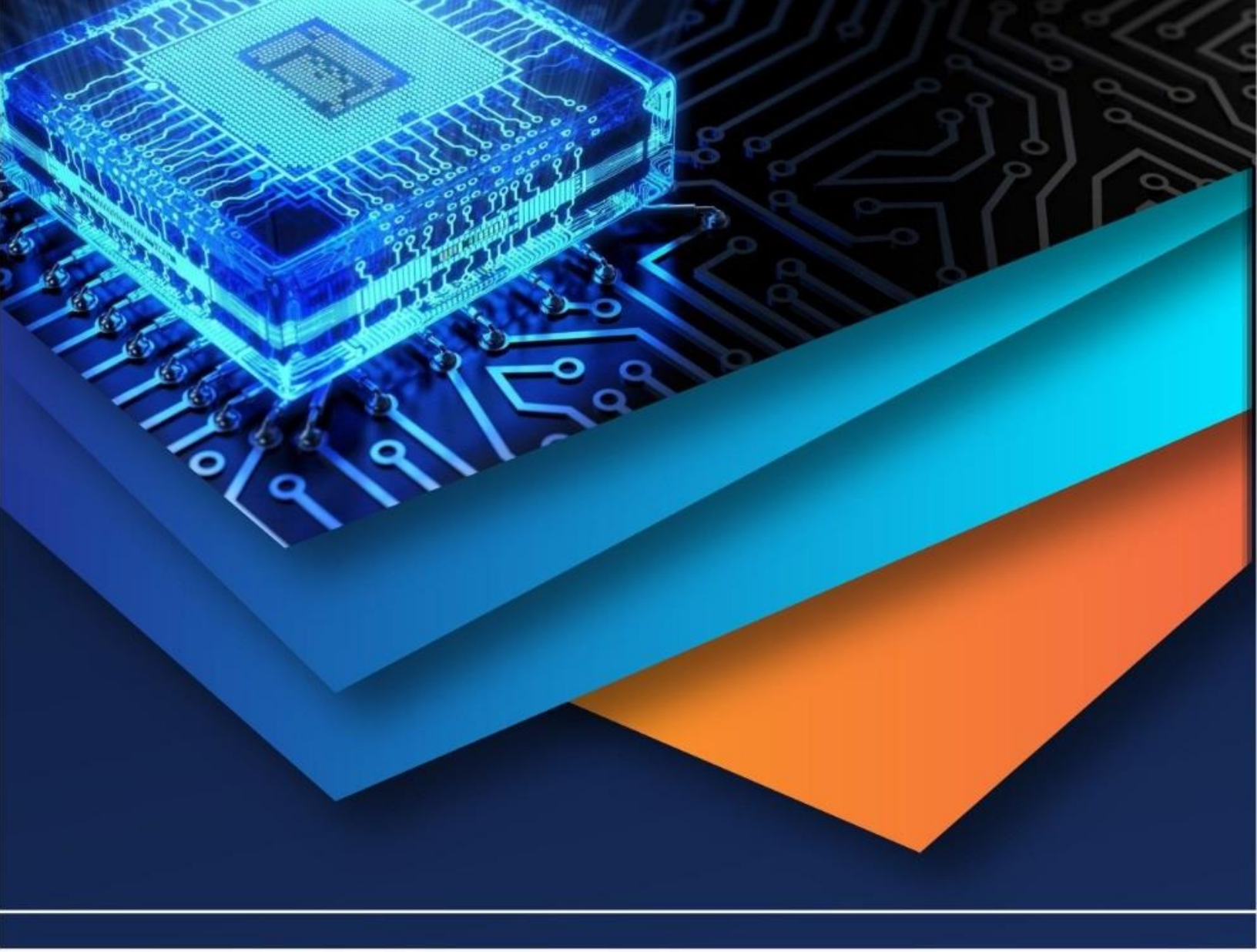
The new method is simple to implement but only applied to numerical data using univariate Gaussian distribution. In future, this study can be further modified to perform over categorical data analysis and use multivariate Gaussian distribution. Since EM is one of the most commonly used clustering tools for data mining problems, the Enhanced EM algorithm may be used for real-world applications like medical imaging, content-based image retrieval, banking, etc. Also, this research work can be enhanced to automatically determine number of clusters and execute the clustering.

VII. ACKNOWLEDGMENT

We thank Alyssa Kate D. Santos and Marvin D. Villegas for sharing their valuable time and effort in the completion of research resources and the development of the EM Clustering Comparison Module. Also, the authors would like to acknowledge the financial support provided by Commission on Higher Education, Kto12 Project Management Unit, Philippines.

REFERENCES

[1] O. Abbas, "Comparisons Between Data Clustering Algorithms," Int. Arab J. Inf. Technol., vol. 5, no. 3, pp. 320–325, 2008.
[2] D. Raja Kishor and N. B. Venkateswarlu, "Hybridization of Expectation-Maximization and K-means Algorithms for Better Clustering Performance," Cybern. Inf. Technol., vol. 16, no. 2, pp. 16–34, 2016.
[3] A. Santos, E. Figueiredo, M. Silva, R. Santos, C. Sales, and J. C. W. A. Costa, "Genetic-based EM algorithm to improve the robustness of Gaussian mixture models for damage detection in bridges," Struct. Control Heal. Monit., vol. 24, no. 3, 2017.
[4] G. Sehgal and K. Garg, "Improved Expectation Maximization Clustering Algorithm," Int. J. Eng. Comput. Sci., vol. 3, no. 5, pp. 6193–6195, 2014.
[5] J. R. Periera, L. A. MArques, and J. M. Da Costa, "An Empirical Comparison of EM Initialization Methods and Model Choice Criteria for Mixtures of Skew-Normal Distributions," Rev. Colomb. Estad., vol. 35, no. 3, pp. 457–478, 2012.
[6] A. O'Hagan and A. White, "Improved model-based clustering performance using Bayesian initialization averaging," Comput. Stat., pp. 1–26, 2015.
[7] X.-L. Meng and D. B. Rubin, "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," Biometrika, vol. 80, no. 2, pp. 267–278, 1993.
[8] J. A. Fessler and A. O. Hero, "Space-Alternating Generalized Expectation Maximisation Algorithm," IEEE Trans. Signal Process., vol. 42, no. 10, pp. 2664–2677, 1994
[9] B. Thiesson, C. Meek, and D. Heckerman, "Accelerating EM for large databases," Mach. Learn., vol. 45, no. 3, pp. 279–299, 2001.
[10] Y. He and C. Liu, "The Dynamic 'Expectation-Conditional Maximization Either' Algorithm," J. R. Stat. Soc., vol. 74, no. 2, pp. 313–336, 2012.
[11] Z. Volkovich, R. Avros, and M. Golani, "On Initialization of the Expectation-Maximization Clustering Algorithm," Glob. J. Technol. Optim., vol. 2, no. June, pp. 117–120, 2011.
[12]]Z. Hu, "Initializing the EM Algorithm for Data Clustering and Sub-population Detection," 2015.
[13] J. Blömer and K. Bujna, "Simple Methods for Initializing the EM Algorithm for Gaussian Mixture Model," in 20th Pacific Asia Conference on Knowledge Discovery and Data Mining, 2013, no. December 2013.
[14] H. Gandomi, X. Yang, S. Talatahari, and A. H. Alavi, "Firefly Algorithm with Chaos," Commun. Nonlinear Sci. Numer. Simul., vol. 18, no. 1, pp. 89–98, 2013.
[15] Fister, X. S. Yang, and J. Brest, "A comprehensive review of firefly algorithms," Swarm Evol. Comput., vol. 13, pp. 34–46, 2013.
[16] X. S. Yang, "Firefly algorithms for multimodal optimization," Stoch. Algorithms Found. Appl., vol. 5792 LNCS, pp. 169–178, 2009.
[17] G. Avendaño-Franco and A. H. Romero, "Firefly algorithm for structural search," J. Chem. Theory Comput., vol. 12, no. 7, pp. 3416–3428, 2016.
[18] S. Yu, S. Yang, and S. Su, "Self-adaptive step firefly algorithm," J. Appl. Math., vol. 2013, no. 1, pp. 1–9, 2013.
[19] A. Sharma and S. Sehgal, "Image Segmentation using Firefly Algorithm," in 2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit, 2016, pp. 99–102.
[20] S. Sundararajan and S. Karthikeyan, "An Efficient Hybrid Approach for Data Clustering Using Dynamic K-Means Algorithm and Firefly Algorithm," J. Eng. Appl. Sci., vol. 9, no. 8, pp. 1348–1353, 2014.
[21] Hassanzadeh and M. R. Meybodi, "A New Hybrid Approach for Data Clustering using Firefly Algorithm and K-means," in 16th CSI International Symposium on Artificial Intelligence and Signal Processing, 2012, pp. 7–11.
[22] X. Han and T. Zhao, "Auto-K Dynamic Clustering Algorithm," J. Anim. Vet. Adv., vol. 4, no. 5, pp. 535–539, 2005.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY