International Journal of Engineering Research and Technology.
ISSN 0974-3154 Volume 11, Number 12 (2018), pp. 1901-1912
© International Research Publication House
http://www.irphouse.com

Jaya: An Evolutionary Optimization Technique for Obtaining the Optimal Dthr Value of Evolving Clustering Method (ECM)

Al-Hakam Ayad Salih[1], Ahmed Hussein Ali[2], Nada Yousif Hashim[3]

[1] Collage of Art, Tikrit University, Salah ad Din Governorate, Iraq.

[2] AL Salam University College, Computer Science Dep. Baghdad, Iraq.

[3] Computer Science Dep, College of Education, Al-Iraqia University Baghdad, Iraq.

ABSTRACT:

Background and Objective: The algorithm of the Evolving Clustering Method (ECM) is an unsupervised clustering technique that provides a fast online and dynamic estimation of clusters. ECM is dynamically estimating the number of clusters in a given set of data points; it is also deployed for finding the existing centers in a given input data space. The number of clusters generated by the ECM is determined by the threshold Dthr value. The bottleneck of the ECM is how to determine the optimal value of Dthr for best clustering processes and prevent the failure of ECM in many datasets. In this paper, we propose employing the victorious and parameter less nature of Jaya optimization technique to obtain the optimal Dthr value of ECM. Materials and Methods: The effectiveness of the proposed JECM algorithm was tested through purity and accuracy on 5 standard datasets hosted by the UCI machine learning database. The experimental database used in this study include Glass, Seeds, Wine, Iris, and Breast Cancer. Results: The experimental results demonstrate that JECM (Jaya- Evolving Clustering Method) yielded a better clustering performance compared to the suitable number of clusters. The JECM offers an improvement in the updating of the ECM's cluster center and radius by considering the accuracy and purity of the outcome as clustering criteria. Conclusion: In this study, a new on-line clustering approach called JECM which was developed from the ECM clustering method.

Keywords: Evolving Clustering Method, Clustering, Machine Learning, Jaya Algorithm, Optimization.

I. INTRODUCTION

Clustering, synonymous to cluster analysis, classification, numerical taxonomy and typological analysis represents the task of grouping object (with descriptive data) such that object in the same group share more similarity when compared to objects in other groups. Any clustering method is either supervised or unsupervised depending on the availability of the truth labels that consist of finite number of discrete assignable classes given some inputs. Most clustering algorithms assume the availability of all data, operate in batch mode and create crisp or fuzzy spheroid shaped clusters. They have been applied in extensively in exploratory data mining and visualization, object detection, novelty detection knowledge discovery and acquisition, statistical data analysis, machine learning, pattern recognition and data representation and compression.

Clustering is a technique in Machine Learning (ML) which involves data points grouping into clusters of similar features[1]. It involves the use of clustering algorithms to classify a given set of data points into specific groups. Theoretically, it is expected that data points in the same cluster ought to share similar features and highly dissimilar to those in the other clusters. Clustering, as an unsupervised learning method[2, 3], is a common statistical data analysis technique deployed in several fields. Clustering is considered as an important unsupervised learning problem[3, 4] as it involves finding the structure of unlabeled data[5, 6]. The focus of both supervised and unsupervised learning is mainly on data analysis[5]; however, reinforcement learning is most suitable for decision-making problems[7, 8]. Online clustering methods have been proposed with single pass operations where proposed models are evolvable both structurally and parametrically. Compared with batch type methods, they are specifically suitable for problems with following characteristics:

a) Stream Type of Data and Very Large Data: Mining and extraction of knowledge from existing dataset that is extremely large and cumulative data streams that, over time, grows significantly in its sheer size poses great challenging to most data modeling methods. For example, the number searches Google performs increased from 60 million during 2000 to 5.7 billion during 2014 on a daily basis.

b) Dynamic Operating Conditions: A common challenge for real-life applications of clustering algorithm has to be with non-stationary data. Facial and voice recognition, metabolic pathways in biological cells, object tracking in computation vision and monitoring of equipment operating under distinct modes under the influence of random noises are just few examples.

c) Uncertainty in Acquired Information: he ability to make subtle adjustment to existing clusters to deal with signals embedded with varying level of drifts imposed by either the physical attributes of the underlying sensing hardware (i.e. GPS signals) along with varying noises types and impacts they have on the fidelity of the signal when interacting with the environment.

d) Building of Intelligent Systems: An intelligent system improves system performance through learning while preserving useful previously gained experience. It is a highly interdisciplinary field involves control theory, artificial and computational intelligence, biologically and life science inspired systems and various engineering disciplines.

One critical element in any clustering method is the distance (similarity) measure employed to determine: 1) Similarity between a new data to existing clusters for data assignment and cluster update purposes and 2) Similarity between existing clusters to evaluate the necessity of structure pruning. 3) In addition, the suitable number of clusters obtained from the clustering algorithm.

Evolving Clustering Method (ECM) is one of the algorithm that have the problem of selecting the optimal number of cluster. The optimal number of clusters in ECM algorithm depend on the threshold value called Dthr value. Evolving clustering methods have advantage to deal with stream type of data when compared to traditional clustering methods that requires all data to be available at the time of operation.

In this paper, the research is focused on employing evolutionary optimization techniques to improve the performance of Evolving Clustering Method (ECM). Jaya used to determine the optimal value of Dthr for best clustering processes and prevent the failure of ECM in many datasets specially with online stream clustering. This work mainly aims to conduct a sensitivity analysis on the influence of Dthr parameters on ECM's clustering outputs. The remaining part of this work is structured thus. Section II presents a description of the ECM algorithm in details, while Jaya evolutionary optimization algorithm is presented in section III. A review of the related literature is provided in section IV while the proposed methodology and datasets description is demonstrated in section V. The experimental setup and results with discussion are presented in section VI while some conclusions drawn from the study and some future works are presented in section VII.

II. EVOLVING CLUSTERING METHOD (ECM)

A. ECM

ECM is a fast clustering method which is based on normalized Euclidean distances (ED)[9] and can be applied two modes (on-line or off-line). ECM has been employed in an on-line mode previously[10, 11] for the prediction of time-series. For the off-line mode, it is an extension of the online mode. In this study, the on-line ECM was deployed to solve the problem of optimal Dthr threshold value selection. The ECM is a fast clustering framework for dynamically estimating the number of clusters in a given set of data points; it is also deployed for finding the existing centers in a given input data space. The ECM is a clustering method that relies on distance[12]. Note that the maximum distance between a given point in any cluster and its associated cluster center is usually less than the user-defined threshold Dthr value. It influences the number of clusters that need to be estimated. If any cluster has a radius which is more than the threshold value (Dthr) during clustering, then, the expansion of such cluster will be suspended. The next section reveals the role of Dthr in achieving a better clustering accuracy.

B. ECM Algorithm

The following steps described the major ECM procedures[13]:

Step 0: The first cluster center (C1) is created by considering the first cluster center (Cc1) as the position of the first example from the input data stream, and its cluster radius (Ru1) value set as 0.

Step 1: Finish the algorithm by processing all the examples of the data stream; otherwise, the current example xi becomes the distance Dij between the considered example and all the cluster centers Ccj already created.

Dij = ||xi – Ccj||, j= 1 to n are estimated.

Step 2: Should a cluster center(s) Ccj exist, for j = 1 to n such that the value of the distance Dij = ||xi- Ccj|| ≤ the radius Ruj, then, the current example xi is assumed to belong to a cluster Cm with the least of these distances: Dim = ||xi – Ccm|| = min (Dij);

where Dij ≤ Ruj, j = 1 to n.

Here, the existing clusters were neither updated nor new ones created, hence, the algorithm reverts to Step 1. However, upon the creation of a new cluster or upon updating an existing cluster, the algorithm will proceed to the next step.

Step 3: From the existing cluster centers, find a cluster with a center Cca and a radius Ru by determining the values Sij = Dij + Ruj, j =1 to n. Then, choose the Cca with the least value Sia: Sia = Dia + Rua = min [14], j =1 to n.

Step 4: Should the Sia be more than 2 * Dthr, then, it implies that xi is not a member of any existing cluster and demands a new cluster to be created as described in Step 0. The algorithm then reverts to Step 1.

Step 5: If Sia is < 2*Dthr, then, the center Cca of cluster Ca is moved by increasing its radius Rua value (updated). The value of the updated radius Ruanew is set to be Sia/2 while the new center Ccanew is placed on the line that links the new input vector xi to the Cca such that the distance between Ccanew and the point xi = Ru. The algorithm reverts to Step 1.

C. Jaya Algorithm

Jaya is a globally recognized simple and robust optimization[15] framework with application to the benchmark function of both unconstraint and constrained problems. Though it is parameterless like the TLBO, it differed from TLBO by not requiring a learner phase (uses only the teacher phase) whereas TLBO uses both learner and teacher phases[16]. Jaya is based on achieving to a given problem by avoiding the worst solution and moving towards the best solution. This algorithm is excellent as it requires only few control parameters such as the number of design variables and the maximum population size/number of generations. There is no need for any specific algorithmic control parameter and does not need extensive parametric tuning before executing computational experiments. The framework of Jaya is outlined in an easy-to-understand manner in Algorithm-1.

Let f(x) represent the objective function (OF) that needs to be minimized. Assume the number of design variables to be ‘n’ (j = 1 to n) and ‘m’ to be the number of individual solutions (k = 1 to m) at any given iteration i.

Furthermore, let f(x) be the candidate that obtains the best solution in a population, while the worst candidate achieves the worst OF value f(x) in the population. If for ith iteration Xj,k,i is the value of the jth variable for the kth candidate, this value Xj,k,i is altered using Eq.1.

X'j,k,i=Xj,k,i+r1,j,i(Xj,best,i-|Xj,k,i|)–r2,j,i(Xj,worst,i-|Xj,k,i|)…(1)

where,

X'j,k,I = the updated new Xj,k,I value, best,I = best Xj,k,I value,

Xj,worst,i = worst Xj,k,I value, r1,j,i and r2,j,i = two random numbers ranging from 0 to 1.

The value of X'j,k,i is considered acceptable if it offers a better value of the OF which will serve as the input in the subsequent iteration. This cycle is continuously repeated until the algorithm is stopped after reaching a predetermined number of iteration (Figure 1).

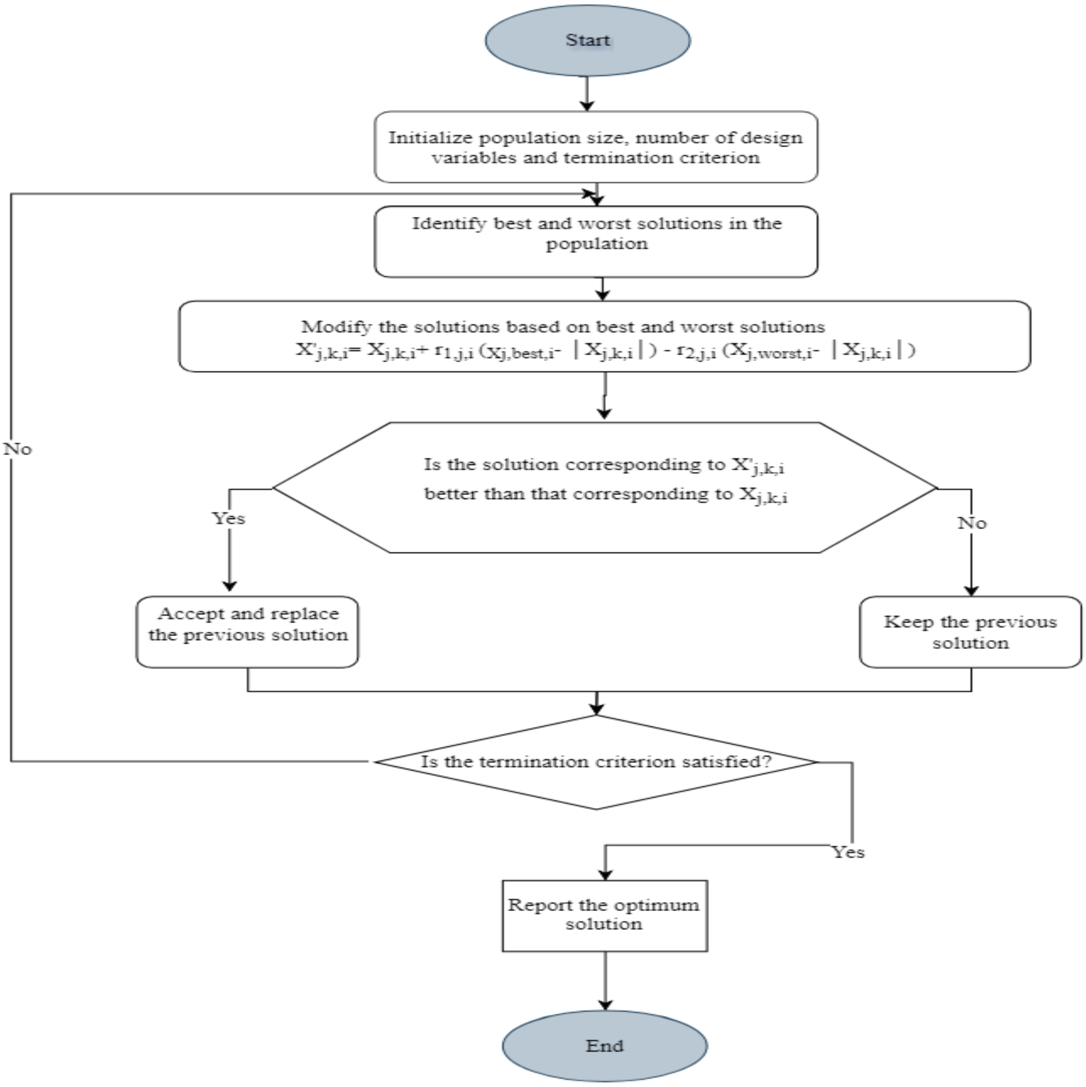
Start
Initialize population size, number of design variables and termination criterion
Identify best and worst solutions in the population
Modify the solutions based on best and worst solutions
$X'_{j,k,i}= X_{j,k,i}+ r_{1,j,i} (X_{j,best,i}- |X_{j,k,i}|) - r_{2,j,i} (X_{j,worst,i}- |X_{j,k,i}|)$
Is the solution corresponding to $X'_{j,k,i}$ better than that corresponding to $X_{j,k,i}$
Yes
No
Accept and replace the previous solution
Keep the previous solution
No
Is the termination criterion satisfied?
Yes
Report the optimum solution
End

Fig.1. Flowchart of the jaya algorithm

III. PREVIOUS WORK

Many research papers have discussed the development of ECM in order to achieve good clustering results. Chandan & Vadlamani[13] performed a sensitivity analysis of the effect of Dthr parameter for ECM over 12 datasets. Sriram et al[17] introduced ECM-ELM for classification by combining ELM with ECM clustering algorithm in order to improve the stability of ELM. Shanmugapriya and Padmavathi[18] archived maximum feature selection by applying Ant Colony Optimization (ACO) and Genetic Algorithm (GA) with ELM. Enrico et al[19] proposed a monitoring model to measure the confidence of decision makers system by employing an auto-associative algorithm based on ECM.

IV. PROPOSED METHOD

The steps of the proposed Jaya Evolving Clustering Method (JECM) are as follows:

Input:

1. Sample data Xi from data streaming without label or target (unsupervised learning).

2. Use Jaya algorithm to obtain the optimal Dthr value.

Output: Clustering results in C.

Step 1: Apply Jaya algorithm to select the best Dthr value of ECM according to the input data.

Step 2: Apply Jaya algorithm on the ECM parameters.

Step 3: Establish the first cluster C1 by taking the first entered data point X1 as the first cluster center (Cc1) with an initial cluster radius (Ru1) value of 0.

Step 4: If all the sample data in the data streaming have been disposed, then, the method is terminated, otherwise, proceed to calculate the distance between current entered data xi and all the existing cluster centers Ccj using the formula Dij = ||xi - Ccj||, j = 1, 2, …, k; where k stands for the number of existing clusters.

Step 5: If Dij ≤ Ruj, j = 1, 2, …, k; this indicates that the current input Dim = xi − Ccm = j =min1 , 2,⋯, k(xi − Cc j data) xi belongs to an existing cluster Cm, and, in this case, xi ∈Cm. if no new cluster is created and no existing cluster center or radius is updated, go back to step 2, otherwise, proceed to the next step

Step 6: Determine the sum of Dij of the current input data xi and all existing cluster centers Cc j and the corresponding cluster radius Ruj, that is, Sij= Dij+ Ruj, j=1, 2, …, k. Select a cluster center Ca (the corresponding cluster center is Cca and the radius is Rua) and make it meet Sia = Dia + Rua = min(Sij), j = 1, 2, …, k.

Step7: If Sia > 2 × Dthr, then, xi does not belong to any existing cluster. Employ the method in Step 1 to create a new cluster, then, go back to Step 2.

Step8: If Sia⩽2×Dthr, then xi∈Ca and Ca is updated by removing its Cca and adding its Rua. The new cluster radius $\mathrm{Ru}_{\mathrm{newa}}$ = Sia/2; remove the new cluster center Cc_{newa} to the ligature of xi and Cca and make it satisfy the condition ||Ccnewa -xi||=Runewa; then, revert to Step 2.

This method is characterized by a good adaptability as it can only deal with the current data when there is a continuous inflow of streaming data. The method is time-efficient as it does not deal with already processed historical data; this makes the method fit better with dynamic streaming data clustering problems compared to the conventional clustering methods. Threshold is used in this method to decide the cluster of the current data; it also uses the least distance to classify the current data.

V. DATA SETS

The effectiveness of the proposed JECM was tested on 5 standard datasets hosted by the UCI machine learning database. The experimental database used in this study include Glass, Seeds, Wine, Iris, and Breast Cancer (Detailed in Table 1). In the table, the first row depicts the data set used while the number of samples, categories, and attributes of the corresponding data sets are shown in the other 3 rows. The employed data sets (Iris, Wine, Seed, and Glass) are free of missing values. Breast Cancer data was extracted from the original dataset of Breast Cancer Wisconsin which has 699 samples, with 16 samples having missing values. The remaining 683 samples do not have missing values but with 10 attributes and 2 actual categories. For an easy assessment of the clustering results, all the datasets in Table 1 are labeled with attribute labels. In each data set, the category attributes are used only to check the purity and accuracy of the process without consulting the clustering process or estimating the value of the clustering results.

Tab.1 Datasets used for the experiments

Dataset	No. of samples	No. of attributes	No. of categories
Glass	214	11	6
Seeds	210	8	3
Wine	178	14	3
Iris	150	5	3
Breast cancer	683	10	2

VI. RESULTS AND DISCUSSION

The sensitivity of the ECM is influenced by the threshold Dthr whose value has a direct effect on the size and number of each cluster in clustering results. Different thresholds and their influence on ECM's clustering results are shown in Figure 2. The investigations in this study are conducted in a PC running on Windows 10 (64bit) with a hardware configuration of Intel i5-2450M CPU 2.5GHz and 10 GB memory. MATLAB R2017a was used to write the methods in a Matlab platform.

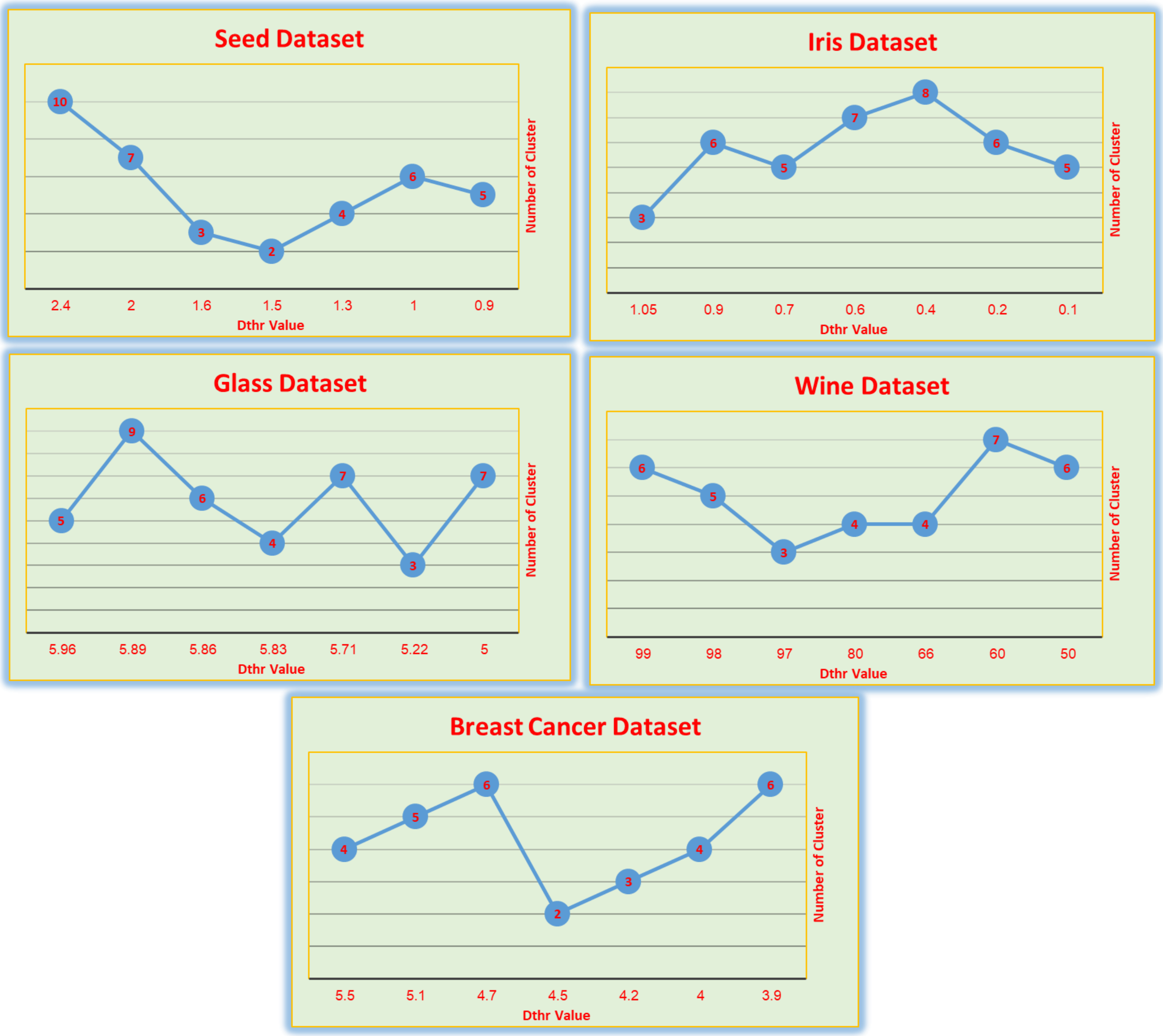

Fig. 2. Influence of optimized Dthr value on the number of clusters for different datasets.

The efficiency of the proposed JECM and that of ECM was tasted in this study based on 3 clustering performance measures, which are the value, the accuracy, and the purity of the objective function. For the sake of convenience, the streaming data were simulated through a row-wise reading of the streaming data in data sets. During clustering analysis, the value of the objective function J (which usually reflects the inter-cluster differences and similarities) often uses the total least squared error which is expressed as:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \left(\| x_j - Cc_i \| \right)$$

where mi = number of samples in the ith cluster. The quality of a clustering result is assessed based on the accuracy rate which is expressed as:

$$accuracy = \frac{\sum_{j=1}^{k} a_j}{N}.$$

Regarding the number of clusters in Table 1, the threshold Dthr is selected to establish the 5 data sets. The value of the objective function, as well as the purity and accuracy of the clustering outcomes after clustering the streaming data, was simulated using ECM and JECM as shown in Table 2.

Tab.2 Comparison of Clustering Results of ECM and JECM.

Dataset	Optimal number of cluster	ECM			JECM		
		Purity	J	Accuracy	Purity	J	Accuracy
Glass	6	0.881	626.5	0.892	0.889	619.4	0.899
Seeds	4	0.843	156.4	0.734	0.897	133.3	0.903
Wine	4	0.788	7144.8	0.661	0.857	5902	0.717
Iris	3	0.841	56.6	0.721	0.903	48.8	0.912
Breast Cancer	3	0.858	2134.7	0.698	0.934	1088	0.937

where N = number of data used during the experiments, and aj = number of the jth cluster in the clustering outcomes which tallied with the actual cluster. The quality of a clustering result can also be checked by the average purity of the clustering results. This criterion is calculated as follows:

$$purity = \frac{1}{k} \sum_{i=1}^{k} \frac{N_i^d}{N_i}.$$

where k = number of clusters, Ni = number of data in the ith cluster, and Ndi = number of data in the main cluster of the ith cluster. Accordingly, many remarkable observation is that the algorithm of JECM alone outperformed the novel method SyncTree presented in[9]. Further, we found that JECM based on jaya optimization achieved better accuracy and purity than the evolving clustering method ECM presented in[11, 20, 21]. on this basis, we deduce that the proposed algorithm can be used as viable alternative for standard ECM and other online clustering method.

VII. CONCLUSION

In this study, a new on-line clustering approach called JECM which was developed from the ECM clustering method was proposed. The JECM offers an improvement in the updating of the ECM's cluster center and radius by considering the accuracy and purity of the outcome as classification criteria. The use of Jaya algorithm in JECM made the value of the objective function of JECM clustering to be better compared to that of the original ECM. There was an improvement in the similarity of clustering results as well. The JECM can be deployed as an independent technique for solving various classification and clustering problems. From the clustering results of using JECM in both on-line and off-line modes, the performance of JECM is comparable to that of other well-known clustering techniques. To further research effort in this direction, the JECM can be improved for on-line adaptive classification and clustering; it can also be employed in other research fields such as mobile robot navigation, online ML for big data analytics, and object tracking.

SIGNIFICANCE STATEMENT

This study aims to improve the clustering process of evolving clustering method (ECM). ECM provides fast online and dynamic estimation of clusters. The optimal number of clusters depend on critical ECM parameter called Dthr value. This study proposed a new method named JECM jaya evolving clustering method. JECM employing the victorious and parameter less nature of jaya optimization algorithm to obtain the optimal Dthr value of ECM. The new method is more suitable in now a day's machine learning applications and fast online prediction.

AUTHOR'S CONTRIBUTION

Ahmed H. Ali conceived of the presented idea. Al-Hakam A. Salih developed the theory and performed the computations. Ahmed H. Ali and Nada Y. Hashim collect the relevant data for this study and verified the analytical methods. Ahmed H. Ali encouraged Al-Hakam A. Salah to investigate the desired results and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Acknowledgements

The authors would like to thank AL-Salam University College, Computer Science Department for their moral support. Special thanks to the anonymous reviewers for their valuable suggestions and constructive comments.

REFERENCES

[1] N. K. Kasabov and Q. Song, "DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction," *IEEE transactions on Fuzzy Systems,* vol. 10, no. 2, pp. 144-154, 2002.

[2] P. Shen, X. Du, and C. Li, "Distributed semi-supervised metric learning," *IEEE Access,* vol. 4, pp. 8558-8571, 2016.

[3] S. A.-b. Salman, A.-H. A. Salih, A. H. Ali, M. K. Khaleel, and M. A. Mohammed, "A New Model for Iris Classification Based on Naïve Bayes Grid Parameters Optimization."

[4] R. A. Hasan, M. A. Mohammed, Z. H. Salih, M. A. B. Ameedeen, N. Țăpuş, and M. N. Mohammed, "HSO: A Hybrid Swarm Optimization Algorithm for Re-Ducing Energy Consumption in the Cloudlets," *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol. 16, no. 5, pp. 2144-2154, 2018.

[5] M. Z. A. Ahmed Hussein Ali, "An Efficient Model for Data Classification Based on SVM Grid Parameter Optimization and PSO Feature Weight Selection," *International Journal of Integrated Engineering,* 2018.

[6] M. Z. A. Ahmed Hussein Ali, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics," *International Journal of Integrated Engineering,* 2018.

[7] R. A. Hasan, M. N. Mohammed, M. A. B. Ameedeen, and E. T. Khalaf, "Dynamic Load Balancing Model Based on Server Status (DLBS) for Green Computing," *Advanced Science Letters,* vol. 24, no. 10, pp. 7777-7782, 2018.

[8] M. A. Mohammed and N. ȚĂPUŞ, "A Novel Approach of Reducing Energy Consumption by Utilizing Enthalpy in Mobile Cloud Computing," *Studies in Informatics and Control,* vol. 26, no. 4, pp. 425-434, 2017.

[9] J. Shao, Y. Tan, L. Gao, Q. Yang, C. Plant, and I. Assent, "Synchronization-based clustering on evolving data stream," *Information Sciences,* 2018.

[10] E. Soares, P. Costa Jr, B. Costa, and D. Leite, "Ensemble of evolving data clouds and fuzzy models for weather time series prediction," *Applied Soft Computing,* vol. 64, pp. 445-453, 2018.

[11] R. Hyde, P. Angelov, and A. R. MacKenzie, "Fully online clustering of evolving data streams into arbitrarily shaped clusters," *Information Sciences,* vol. 382, pp. 96-114, 2017.

[12] Q. Song and N. Kasabov, "ECM-A novel on-line, evolving clustering method and its applications," *Foundations of cognitive science,* pp. 631-682, 2001.

[13] C. Gautam and V. Ravi, "Evolving clustering based data imputation," in *Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on,* 2014, pp. 1763-1769: IEEE.

[14] K. Senathipathi and K. Batri, "An analysis of particle swarm optimization and genetic algorithm with respect to keystroke dynamics," in *Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on,* 2014, pp. 1-11: IEEE.

[15] R. Rao, "Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems," *International Journal of Industrial Engineering Computations,* vol. 7, no. 1, pp. 19-34, 2016.

[16] H. M. Pandey, "Jaya a novel optimization algorithm: What, how and why?," in *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, 2016, pp. 728-730: IEEE.

[17] S. Ravindran, C. Gautam, and A. Tiwari, "Keystroke user recognition through extreme learning machine and evolving cluster method," in *Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on*, 2015, pp. 1-5: IEEE.

[18] D. Shanmugapriya and G. Padmavathi, "An efficient feature selection technique for user authentication using keystroke dynamics," *IJCSNS International Journal of Computer Science and Network Security,* vol. 11, no. 10, pp. 191-195, 2011.

[19] E. Zio, P. Baraldi, and W. Zhao, "Confidence in signal reconstruction by the Evolving Clustering Method," in *Prognostics and System Health Management Conference (PHM-Shenzhen), 2011*, 2011, pp. 1-7: IEEE.

[20] D. G. Márquez, A. Otero, P. Félix, and C. A. García, "A novel and simple strategy for evolving prototype based clustering," *Pattern Recognition,* vol. 82, pp. 16-30, 2018.

[21] C. G. Bezerra, B. S. J. Costa, L. A. Guedes, and P. P. Angelov, "A new evolving clustering algorithm for online data streams," in *Evolving and Adaptive Intelligent Systems (EAIS), 2016 IEEE Conference on*, 2016, pp. 162-168: IEEE.