DOI: http://dx.doi.org/10.26483/ijarcs.v15i2.7071

ISSN No. 0976-5697

Volume 15, No. 2, March-April 2024

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A New Hybrid Optimization Algorithm to Solve the Clustering Problem

Subash Kumar[1,*] , Sikander Singh Cheema[1]
[1,*]Department of Computer Science and Engineering,
Punjabi University, Patiala,India

Abstract: This paper presents a novel hybrid technique, merging the K-means algorithm with Genetic Algorithm (GA), aiming to enhance clustering performance. This approach leverages the strengths of both algorithms, enabling improved cluster generation by overcoming individual algorithmic limitations. The GA-KM algorithm is introduced to aid K-means in avoiding local optima, with GA exhibiting proficiency in determining optimal cluster initialization and parameter optimization. The focus is on developing a GA-based algorithm for generating high-quality clusters efficiently. Notably, the research explores the application of this hybrid approach to address issues in the educational domain, specifically for out-of-school children. The fitness function in GA is tailored to the problem area, emphasizing the need for an appropriate system to study and address school children's problems. The research proposes a hybrid algorithm (KM-GA-NM-PSO) that amalgamates the best features of existing algorithms, thereby overcoming individual limitations and promising superior results. This hybridization is expected to yield high-quality clusters with minimal function evaluations, outperforming other methods by producing clusters with small standard deviations on selected datasets. The proposed approach, combining KM, GA, NM, and PSO algorithms, demonstrates improved data clustering quality and algorithmic efficiency.

Keywords*:* Hybrid clustering, K-means, Genetic Algorithm, Educational data set, Cluster quality

1. Introduction

An improvement over the algorithm is a hybrid technique based on combining the K-means algorithm with various other algorithms. The combined approach of different algorithms therefore provides better performance using the goodness of the whole algorithm to overcome the disadvantage of any particular algorithm. Genetic algorithm is one of the most commonly used evolutionary algorithm techniques to solve a clustering problem. Therefore, a hybrid data clustering algorithm based on GA and k-means (GA-KM) [1][2], which uses the advantages of both algorithms. The GA-KM algorithm helps the k-means algorithm to escape local optimum. GA has been shown to be able to determine the best cluster initialization and to optimize initial parameters [3][4][5]. GA defines a randomly generated population of people. These people are involved in the generation of new and better offspring by mutation / crossover. Decision on better offspring / individuals is achieved by fitness. The greatest benefit of genetic algorithms is that the fitness function can be changed to change the algorithm's behaviour. There is a wide variety of representations of individual or chromosomes [6][7][8]. The solutions are traditionally represented using fixed length strings, in particular binary strings, but alternative encoding has been developed. The main focus of the GA-based algorithm was to generate high-quality clusters in optimized time [9][10][11]. The focus of the current research was to use GA as an initial centroid selection tool and to study the performance of improved clustering of k-means. The applications of GA-based k means have been tested in literature on standard data sets, but educational data set specifically from the problem of school children has not been investigated. Current research has focused on developing an appropriate system to study school children's problems using basic k-means and improved k-means (GA with k-means). Consequently, the approach to the development of a new algorithm was problematic and the selection criteria or initial centroid influenced the nature of the domain [12][13][14]. In short, according to the problem area, the fitness function in GA has been defined. Apart from identifying preferable technique for out of school children problem, there is always a need to analyze quality of clusters. There will be good method to measure the quality of the better clusters and performance of clustering. The hybrid (KM-GA-NM-PSO). Algorithms contain all the best features of the existing algorithm that overcome the limitations of the individual algorithm when combined. The improvement of this combined approach will lead to even better results. This will be requiring a minimum number of evaluations of functions to achieve the optimum solution. Compared to other methods, the hybrid approach will be produces high-quality clusters with small standard deviations on selected data sets. It is proposed to combine with KM, GA, NM, PSO algorithms. This combination of hybrids improves the quality of data clustering and improves the algorithm [15][16][17].

2. Experimental results

Step 1: K-mean method applies randomly choose k centroids from dataset for desired clusters assign to each data object to the cluster with the closet centroids [18][19][20]. Update the centroids by calculating the mean value of object within clusters. Repeat step 1.2 and 1.3 until termination centroids are met.

Step 2: Generate initial population of size i({j1, j2,j3,.....,ji}).

- J1= k-mean (dataset)
- J2=min (dataset)
- J3=mean (dataset)
- J4=max (dataset)
- J5=Ji= random value of (dataset)

Step3: GA algorithm apply

- Apply crossover operator on N particle (GA) [21][22].
- Apply mutation operator on update N particle (GA).

Step4: NM simplex method apply

- Initialization: Generate a population of size 3N+1.
- Evaluation and Ranking: Evaluate the fitness of each particle rank them on the basis of fitness.
- Apply NM operator to the top N+1 particle and replace the (N+1) particle with the update.

Step PSO algorithm apply

- Apply PSO operator for updating the remaining 2N particles.
- Selection: from the population select the global best particle and the neighbourhood best particles. Velocity Update: apply update to the 2N particle with worst fitness according equations (3) & (4);

Step5: If the termination conditions are not meet then go to back 4. 2.

3. Simulations

Iris Data Set [23][24][25] we used the Iris data set to bring our algorithms a pragmatic result. In this case, each data set in the Iris Data Set has the number of their own distributions that these items of clusters and data are important to. Iris is used to set up a good comparison and algorithm for data sets. In this data set (n=150, d=4, k=3) it has three equal squares of 50 squares. In this data set we have 150 samples. It covers each class type of a class iris Flowers, in which four-digit properties are also included. These data sets are such that the length of the sepal in cm, width and height of the petals Widths are in cent-meters. There is no missing value in this data set [26][27][28].

4. Performance measure

The Iris data set has been used in separate different algorithms, a predominantly KM algorithm, GA, NM,PSO Algorithm and K-GA-NM-PSO Algorithm have been developed in a table. In which good results have been found and the individual's best performance has been received. That compares to other clustering algorithms. K-mean algorithm in some cases there are problems. Just as in the beginning, there may be a set of solutions for the K-GA matching solution to the problem of a satellite base and its solutions. So, we are using the PSO algorithm. With the help of algorithms [29][30][31], it helps to maintain the integrity of all algorithms and simultaneously solve their problems. This is how the NM algorithm has been defeated again. NM algorithm helps us to provide a lot of efficient local research process from algorithms. But the NM algorithm is dependent on the starting point and this convergence is sensitive to choose the randomly the starting point and this can also be algorithms increase percentage in algorithm [32][33][34][35].

Table 1: Results

K Value	k-mean	GA	NM	PSO	k-mean+ GA+NM+PSO
K=1	68.6166	66.0783	60.0123	70.2568	35.1443
K=2	82.6219	68.635	59.3256	94.2564	50.6701
K=3	129.5325	92.3585	91.6584	135.2567	20.3042
K=4	203.5256	150.1065	149.2569	278.2547	48.0000
K=5	355.2576	121.4141	360.5698	396.4567	91.2340
K=6	328.016	198.7141	365.1245	421.2584	68.2677
K=7	432.2051	268.4564	456.2584	547.1234	48.8133
K=8	516.0121	339.368	591.4568	621.5487	16.2100
K=9	645.2582	367.8258	679.2465	754.2547	54.5613
K=10	766.1073	241.8844	790.4658	875.2547	33.0444

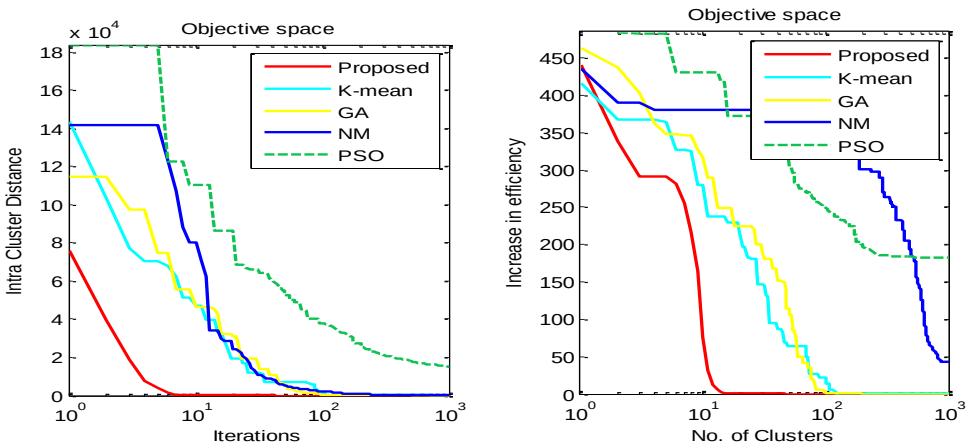

Fig 1. Comparison Results

The comparison performance shown in the table is making it show as KM, GA, NM, PSO vs. k-mean-GA-NM-PSO people have been reproduced and individual clusters are made in and between them. And calculation of performance details etc. Thus all the sets of KM-GA-NM-PSO algorithms are tested and as well as solutions of high-quality cluster have been developed. Which are designed in the form of distance of the best inter cluster. Also discovered are the storms standard deviation and the smallest found to near optimal solution of the run other algorithm may trap local optima in some of run. It is found to better results , thus KM-GA-NM-PSO keeps Algorithm a stronger one. This K-MEAN algorithm requires a smaller number compared to other algorithms and it is in relation to the functional visits. In this way we can say that by using the result of K-MEAN in KM-GA-NM-PSO, the GA is in a good way, which is a great way to get access to a great tool from a single GA Is of algorithm produces new generation population from traffic for generation of pig production and the environment is resolved to a new baby environment. In this way a child's solution has many features of his measurement which can be created from new parents to newborn babies. But still there is not a good start with G.A., a good start with the combination of KM-GA to overcome its shortage can be started and new parents from new parents can be produced, and a suitable population size can also be made. Thus, KM-GA can be better equipped with algorithmic combination than PSO, meaning that the new population can be created at the onset of the cluttering process and can be speeded up in this situation and the health status can be discarded because it Less cluttering needs lesser working people, After we have done all the procedure, we can say that the outcome of PSO and NM-PSO clustering can be revised. With the K-MEAN algorithm, this hybrid algorithm ends with the first K-MEAN algorithm and if there is no change in this cluster's satire rayon vector, in the case of K-PSO, K-MEAN algorithm results in one particle used in the form. The 5N-1 particles start randomly, so this hybrid is used in K-GA-NM-PSO. The 3N-1 angle creates the points continuously and NM-PSO then forms this form to complete the process.

5. Conclusion

The proposed hybrid clustering algorithm, combining K-means with Genetic Algorithm (GA), presents a promising approach for addressing clustering challenges. The GA-KM algorithm effectively mitigates local optima issues, optimizing cluster initialization and parameters. The research focuses on the unique application of this hybrid approach to educational data sets, particularly addressing the problem of out-of-school children. The tailored fitness function in GA proves crucial for adapting the algorithm to the specific problem area. The proposed hybrid algorithm, incorporating KM, GA, NM, and PSO, demonstrates superior performance, producing high-quality clusters with minimal function evaluations and small standard deviations. This research not only advances clustering techniques but also addresses real-world educational challenges, showcasing the potential of hybrid algorithms in improving both algorithmic efficiency and data clustering quality.

6. References

[1] Garg, R. K., Soni, S. K., Vimal, S., & Dhiman, G. (2023). 3-D spatial correlation model for reducing the transmitting nodes in densely deployed WSN. Microprocessors and Microsystems, 103, 104963.

[2] Dehghani, M., Bektemyssova, G., Montazeri, Z., Shaikemelev, G., Malik, O. P., & Dhiman, G. (2023). Lyrebird Optimization Algorithm: A New Bio-Inspired Metaheuristic Algorithm for Solving Optimization Problems. Biomimetics, 8(6), 507.

[3] Mekala, M. S., Dhiman, G., Park, J. H., Jung, H. Y., & Viriyasitavat, W. (2023). ASXC $^{2} $ Approach: A Service-X Cost Optimization Strategy Based on Edge Orchestration for IIoT. IEEE Transactions on Industrial Informatics.

[4] Kumar, A., Misra, R., Singh, T. N., & Dhiman, G. (2023). APO-AN feature selection based Glorot Init Optimal TransCNN landslide detection from multi source satellite imagery. Multimedia Tools and Applications, 1-38.

[5] Rajinikanth, V., Razmjooy, N., Jamshidpour, E., Ghadimi, N., Dhiman, G., & Razmjooy, S. (2023). Technical and Economic Evaluation of the Optimal Placement of Fuel Cells in the Distribution System of Petrochemical Industries Based on Improved Firefly Algorithm. In Metaheuristics and Optimization in Computer and Electrical Engineering: Volume 2: Hybrid and Improved Algorithms (pp. 165-197). Cham: Springer International Publishing.

[6] Alferaidi, A., Yadav, K., Yasmeen, S., Alharbi, Y., Viriyasitavat, W., Dhiman, G., & Kaur, A. (2023). Node Multi-Attribute Network Community Healthcare Detection Based on Graphical Matrix Factorization. Journal of Circuits, Systems and Computers, 2450080.

[7] Singh, S. P., Dhiman, G., Juneja, S., Viriyasitavat, W., Singal, G., Kumar, N., & Johri, P. (2023). A New QoS Optimization in IoT-Smart Agriculture Using Rapid Adaption Based Nature-Inspired Approach. IEEE Internet of Things Journal.

[8] Singh, S. P., Piras, G., Viriyasitavat, W., Kariri, E., Yadav, K., Dhiman, G., ... & Khan, S. B. (2023). Cyber Security and 5G-assisted Industrial Internet of Things using Novel Artificial Adaption based Evolutionary Algorithm. Mobile Networks and Applications, 1-17.

[9] Khan, M., Kumar, R., Aledaily, A. N., Kariri, E., Viriyasitavat, W., Yadav, K., ... & Vimal, S. (2023). A Systematic Survey on Implementation of Fuzzy Regression Models for Real Life Applications. Archives of Computational Methods in Engineering, 1-21.

[10] Gulia, P., Kumar, R., Viriyasitavat, W., Aledaily, A. N., Yadav, K., Kaur, A., & Dhiman, G. (2023). A Systematic Review on Fuzzy-Based Multi-objective Linear programming Methodologies: Concepts, Challenges and Applications. Archives of Computational Methods in Engineering, 1-40.

[11] Yadav, A. P., Davuluri, S. K., Charan, P., Keshta, I., Gavilán, J. C. O., & Dhiman, G. (2023, February). Probabilistic Scheme for Intelligent Jammer Localization for Wireless Sensor Networks. In International Conference on Intelligent Computing and Networking (pp. 453-463). Singapore: Springer Nature Singapore.

[12] Athawale, S. V., Soni, M., Murthy, K., Dhiman, G., & Singh, P. P. (2023, February). Weakly Supervised Learning Model for Clustering and Segmentation of 3D Point on Cloud Shape Data. In International Conference on Intelligent Computing and Networking (pp. 531-543). Singapore: Springer Nature Singapore.

[13] Pande, S. D., Kumaresan, T., Lanke, G. R., Degadwala, S., Dhiman, G., & Soni, M. (2023, February). Bidirectional Attention Mechanism-Based Deep Learning Model for Text Classification Under Natural Language Processing. In International Conference on Intelligent Computing and Networking (pp. 465-473). Singapore: Springer Nature Singapore.

[14] Singh, N., Virmani, D., Dhiman, G., & Vimal, S. (2023). Multi to binary class size based imbalance handling technique in wireless sensor networks. International Journal of Nanotechnology, 20(5-10), 477-511.

[15] Yadav, K., Al-Dhlan, K. A., Alreshidi, H. A., Dhiman, G., Viriyasitavat, W. G., Almankory, A. Z., ... & Rajinikanth, V. A novel coarse-to-fine computational method for three-dimensional landmark detection to perform hard-tissue cephalometric analysis. Expert Systems, e13365.

[16] Jiby, B. J., Sakhare, S., Kaur, M., & Dhiman, G. (2022). Multi-Criteria Decision Making in Healthcare: A Bibliometric Review. Demystifying Federated Learning for Blockchain and Industrial Internet of Things, 186-213.

[17] S. Jung, Queen-bee evolution for genetic algorithms, Electronics Letters 39 (2003) 575–576.

[18] A. Karci, Imitation of bee reproduction as a crossover operator in genetic algorithms, in: PRICAI, 2004, pp. 1015–1016.

[19] H.F. Wedde, M. Farooq, Y. Zhang, BeeHive: An e cient fault-tolerant routing algorithm inspired by honey bee behavior, in: M. Dorigo, M. Birattari, C. Blum, L.M. Gambardella, F. Mondada, T. Stutzle¨ (Eds.), Ant Colony Optimization and Swarm Intelligence, 4th International Workshop, ANTS 2004, Brussels, Belgium, September 5 - 8, 2004, Proceedings, number 3172 in Lecture Notes in Computer Science, Springer, 2004, pp. 83–94.

[20] N. Gordon, I.A. Wagner, A.M. Brucks, Discrete bee dance algorithms for pattern formation on a grid, in: Proceedings of the IEEE/WIC International Conference on Intelligent Agent Technology, IAT '03, IEEE Computer Society, Washington, DC, USA, 2003, pp. 545–549.

[21] M. Chau, R Cheng., & B. Kao (2005, December) Uncertain data mining: A new research direction. In Proceedings of the Workshop on the Sciences ofthe Artificial, Hualien, Taiwan (pp. 199-204).Applications, 2010. 37(7): p. 4966-4973.

[22] J. Barker, (1958). Simulation of Genetic Systems by Automatic Digital Computers. Australian Journal of Biological Sciences, 11(4), 603-612.

[23] H.J. Bremermann (1958). The evolution of intelligence: The nervous system as a model of its environment: University of Washington, Department of Mathematics.

[24] H. K Bansal, Simulation of Genetic Algorithm Processor, International Journal of Application or Innovation in Engineering and Management IJAIEM, 2012.

[25] P.G. Gonnade, & S. Bodkhe (2012). Geneticalgorithm for task scheduling in distributed heterogeneous system. International Journal ofAdvanced Research in Computer Science and Software Engineering, 2(10).

[26] K.N. Sastry (2007). Genetic algorithms and geneticprogramming for multiscale modeling: Applications in materials science and chemistry and advances in scalability: ProQuest.

[27]W. Spendley, Hext,G. R., & F.R.Himsworth (1962). Sequential application of simplex designs in optimization and evolutionary operation technometrics, 4, 441–461.

[28]J.A.Nelder, & R. Mead (1965) A simplex method for function minimization. Computer Journal, 7, 308–313.

[29]D.M.Olsson, & L.S. Nelson, (1975).The Nelder–Mead simplex procedure for function minimization. Technometrics, 17, 45–51.

[30] J.Kennedy, and R. C. Eberhart : Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks (1995) 1942-1948.

[31] R. C. Eberhart and Y. Shi.: Comparison between genetic algorithms and particle swarm optimization. In: Proceedings of the 7th Annual Conference on Evolutionary Programming (1998)

[32]J. Kennedy and R. C. Eberhart, Swarm intelligence. San Mateo: Morgan Kaufmann, 2001.69-73

[33] K.E. Parsopoulos, Particle Swarm Optimization and Intelligence: Advances and applications .Hershey, PA, USA: IGIGlobal,2010.

[34]R.C.Eberhart& Shi, Y. (2001). Tracking and optimizing dynamic systems with particle swarms. In Proceedings of the Congress on Evolutionary Computation, Seoul, Korea (pp. 9ems)

[35]Hu, X., & Eberhart, R. C.(2001). Tracking dynamic systems with PSO: where's the cheese? In Proceedings of the Workshop on Particle Swarm Optimization, Indianapolis, IN, USA.