IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Hybrid Optimized Algorithms for Solving Clustering Problems in Data Mining

S. Karthikeyan[1*], A. Dhakshina Moorthy[2]

[1]Dept. of Computer Science, Rathinam College of Arts and Science, Coimbatore, Tamilnadu, India
[2]Freelancer – Data Scientist

[*]*Corresponding Author: s.karthics@gmail.com, Tel.: +91 9578682007*

DOI: https://doi.org/10.26438/ijcse/v7i4.928932 | Available online at: www.ijcseonline.org

Accepted: 19/Apr/2019, Published: 30/Apr/2019

***Abstract*—** In this paper, Cluster analysis is a group objects like observations, events etc based on the information that are found in the data describing the objects or their relations. The main goal of the clustering is that the objects in a group will be similar or related to one other and different from (or unrelated to) the objects in other groups. Extracting relevant information from large database is attaining huge significance. Clustering of relevant information from large database becomes difficult. The major objective of this work is to proposed novel clustering methods for solving clustering problem. It is used to separate the data set into a significant set of reciprocally limited clusters with respect to relationship of data and it is used to create the more number of data in the same manner surrounded by a group and extra various among groups. Data clustering is a vital concept of mining as it partitions the given dataset into meaningful set of clusters based on data similarity. This concept enhances the computation efficiency in the data analysis processes

***Keywords*—**Clustering, ABC Algorithm, PSO and FA Algorithm, MOSSSA-HAC, MOSSCS-MHAC Algorithms

I. INTRODUCTION

Clustering is an information mining strategy to aggregate the comparative information into a cluster and disparate information into various groups. The objective of clustering is to gather information into groups to such an extent that the similitude between information individuals inside a similar cluster are maximal while likenesses between the information individuals from various clusters are negligible. The optimization performance is not achieved and hence the overall clustering performance is reduced significantly. To overcome the above mentioned issues, the optimization based clustering algorithms are proposed. The optimization algorithms are such as Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) algorithm, Firefly Algorithm (FA) and Scatter Search (SS) used for improving the optimal solutions for the specified dataset.

II. LITERATURE SURVEY

In this section, the advantages and disadvantages of existing methods are discussed to lead the proposed research in better way.

Gan *et al* (2009), presented the genetic fuzzy k-Modes algorithm for clustering the categorical collections. To accelerate the joining procedure of the algorithm, it utilized the one- step fluffy k-Modes algorithm. Be that as it may it has issue along with taking care of in blended information qualities. Abraham *et al* (2007), presented a strategy for clustering complex and linearly non-separable datasets, with no earlier information of the quantity of normally happening clusters. It enhances the power and precision of the outcomes. Yunfeng Xu, *et al* (2013), recommended a New Artificial Bee Colony (NABC) algorithm, which adjusts the search example of both employed and onlooker bees. Be that as it may it has issue alongside premature convergence in the later search period.

Hassanzadeh *et al* (2012), discussed firefly algorithm which is a swarm based algorithm that is used for solving optimization problems. It increases the accuracy and capability of clustering higher but it has issue with time complexity. Xiaohui Yan *et al* (2012) introduce a Hybrid Artificial Bee Colony (HABC) algorithm for information clustering. The motivator instrument of HABC is upgrading the data trade between honey bees by presenting the crossover operator of GA to ABC. The HABC algorithm is then employed for information clustering. Pacheco *et al* (2005) presents metaheuristic technique in view of the SS approach for the non-hierarchical clustering issue under the basis of minimum Sum-of-Squares clustering. This algorithm fuses systems in light of various procedures, for example,

local search and path relinking. The point is to acquire quality arrangements with short calculation times.

Sangeetha *et al* (2017) presented Similarity based Clustering (ISC) algorithm to provide an efficient big data opinion mining. It exploits the adjacency matrix and merges the clusters into a single cluster. It improves time consumption, memory utility and accuracy. Tanir *et al* (2017) used k-means method to find the initial cluster centers for gene dataset. It shows that the higher accuracy for given dataset.

III. MOTIVATION AND SCOPE OF THE RESEARCH

This research has been motivated by the huge body of literature existing on classification and clustering approaches and their limitations. The research reveals that the conventional techniques of clustering and classifications, though the methods are able to tackle the problems sometimes unable to achieve optimal results by means of accuracy and unable to extend models that are comprehensible in nature. While managing high dimensional information, conventional clustering algorithms that register the separation in full dimensional space flop in distinguishing concealed connections of the fundamental structure because of the reason that the closest neighbour of an example might be almost as inaccessible as the most farthest neighbour. To adapt to the issue of high dimensional feature spaces, feature reduction and feature selection strategies have been utilized. In any case, these procedures don't bargain successfully with clusters and doesn't guarantee optimal solutions for different datasets. Consequently, there is a requirement for more generalized techniques procedures that can be utilized to get important clusters in fluctuating datasets.

IV.LIMITATIONS OF EXISTING CLUSTERING ALGORITHM

In realistic situations, the clusters formed may not be in accordance with the domain expert. Numerous researchers have been tried several methods to lessen such difficulties along with limited success. Computational complexity is still an issue in existing research and it does not handle the imbalanced dataset more effectively. To overcome the above mentioned issues, the hybrid techniques can be used to obtain optimal solutions through developing best clusters in further research.

V. OBJECTIVES

The main targets of this research are as follows:

- To improve the data clustering accuracy by using hybrid techniques for larger dataset.
- To achieve the optimal clustering performance by finding the best fitness function.
- To develop hybridized optimization based clustering techniques for obtaining high quality cluster results.
- To improve the classification accuracy and time complexity by selecting the optimal features from scalable dataset.

VI. RESEARCH CONTRIBUTION

This research is concerned with analyzing and developing optimal and hybridized clustering methods for achieving the high quality clustering performance. The proposed techniques are focused to reduce the misclassification error and computational complexity.

A Hybrid Clustering Approach using Artificial Bee Colony (ABC) and Particle Swarm Optimization (PSO)

In this examination work, joined type of PSO algorithm and ABC algorithm (PSABC) is utilized for clustering reason. Clustering is the way toward perceiving characteristic groupings or clusters in multidimensional information in view of some closeness measures. In PSO, each bird inspects the search space from its new local area, and the procedure rehashes until the point when the run touches base at a favored destination. Presently assesses the molecule's fitness assessment with its past pbest. The best current fitness assessment with the population gbest values is thought about. On the off chance that the present esteem is superior to the population gbest, at that point reset the gbest to the present best position and the fitness esteem to current fitness esteem.

In ABC algorithm, the arrangement of the optimization issue is spoken to by the area of a food source and the nature of the arrangement is spoken to by the nectar measure of the source (fitness). By utilizing target work, the best arrangements were acquired. In this strategy for hybridization, the optimal estimations of people produced by the ABC are provided to the PSO as its beginning stage. Customarily the PSO randomly creates its first individual sets, however for this situation of hybridization that is dealt with by giving the beginning stage to the PSO and it is the last esteems for people produced by the ABC. The local search execution of ABC algorithm relies upon neighborhood search and greedy selection instruments performed by employed and onlooker bees. The global search performance of the algorithm relies upon random search process performed by scouts and neighbor arrangement generation system performed by employed and onlooker bees.

An Efficient Clustering Approach Using Hybrid Swarm Intelligence Based Artificial Bee Colony- Firefly Algorithm

This examination work presents probability of a novel approach Hybrid Artificial Bee Colony-Firefly Algorithm (HABC-FA) for clustering to take care of the clustering issue in the benchmark datasets like Fisher's iris dataset. The pre-processing system incorporates evacuation of repetitive and

irrelevant features from the datasets. This work utilized Independent Component Analysis (ICA) for expulsion of irrelevant features from the dataset. In feature selection process, Genetic Algorithm (GA) is connected which is utilized for distinguishing a reasonable subset of the most helpful features in the iris dataset. While GA is moderately negligent to unnecessary features in the dataset, they appear to be a phenomenal decision for the premise of a further robust feature determination procedure for enhancing the execution of clustering process. In clustering process, swarm intelligence based HABC-FA is utilized to locate the optimal solution. In the ABC algorithm, the main portion of the colony comprises of the employed artificial bees and the second half incorporates the onlookers. For each food source, there is just a single employed bee. At that point enforce FA which plans this genuine flashing behaviour and utilized this behaviour in the onlooker stage bee's genome with similar informational collection objective function of the clustering issue to be enhanced with HABC algorithm. The proposed technique utilizing the HABC-FA algorithm in the clustering system, the objective of every individual bee in ABC is assumed as information purposes of the iris dataset which is utilized to create the best clustering arrangement. To deliver a decent solution vector, every one of the information point gatherings of the ABC algorithm must participate with the FA and the data from every one of the information purpose of the attractive swarm (i.e. fireflies) should be utilized which gives the best separation esteem between the information purposes of the benchmark iris dataset.

Optimal Clustering Approach of Multi-Objective Scatter Search Simulated Annealing with Hierarchical Agglomerative Clustering for Clustering Problems

In the third work, the Multi-Objective Scatter Search Simulated Annealing with Hierarchical Agglomerative Clustering (MOSSSA-HAC) approach is proposed. The preprocessing is done by Modified K-Nearest Neighbour (MkNN) based method which selects the attributes with similar to the attribute of interest to impute missing values. At that point play out the cluster based under-sampling to cluster entire training samples into K groups at that point pick fitting training samples from the determined clusters. To accomplish the multi objective feature selection, MOSSSA is connected in this exploration. SA is utilized to acquire the global optimum when the temperature is diminished vastly moderate.

Hierarchical Agglomerative Clustering (HAC) is used to improve the clustering accuracy by using more informative data. Agglomerative clustering begins with N clusters, each of which incorporates precisely one information point. A progression of merge operations at that point took after, that inevitably powers all articles into a similar gathering. Agglomerative algorithms start with every component as a different cluster and merge them into progressively bigger clusters.

Multi-Objective Scatter Search Using Cuckoo Search with Modified Hierarchical Agglomerative Clustering for Resolving Clustering Problems

In the fourth work, Multi-Objective Scatter Search with Cuckoo Search algorithm with Modified Hierarchical Agglomerative Clustering (MOSSCS-MHAC) is proposed for resolving the convergence problem. In the initial stage, the Modified KNN is used for pre-processing followed by the under sampling process for error reduction. Then the multi-objective feature selection is performed using MOSSCS while the final clustering is achieved using MHAC algorithm. This approach provides optimal clustering with high accuracy. The multi-objective feature selection is performed using MOSSCS while the final clustering is achieved using MHAC algorithm. This approach provides optimal clustering with high accuracy. The experimental analysis provides that the suggested MOSSCS-MHAC gives high values of precision, recall and f-measure than the current algorithms.

VII. RESULTS AND DISCUSSION

The performance evaluation of the work is done by comparing the proposed work with the existing algorithm based on some parameters. The suggested method has been computed through three data sets from various knowledge fields for the reason of computing the performance and effectiveness of proposed MOSSSA-HAC. The suggested MOSSSA-HAC, Hybrid ABC-FA algorithm and PSABC algorithms were distinguished with respect to recall, precision and f-measure metrics. The data sets were available from UCI Machine Learning Repository.

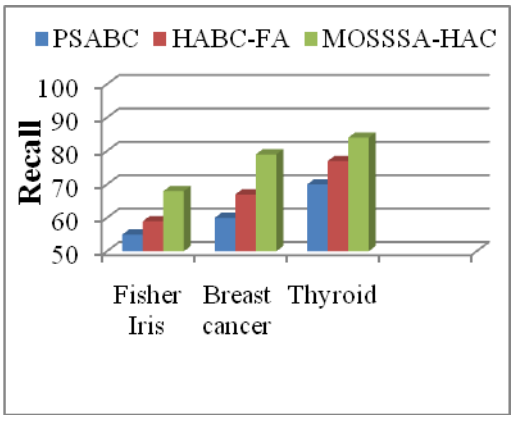

Figure 1 Recall

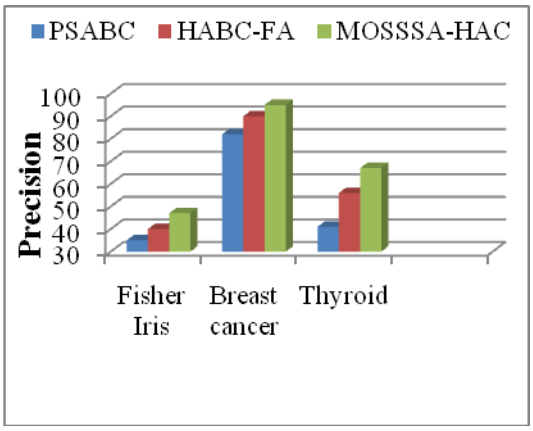

Figure 2 Precision

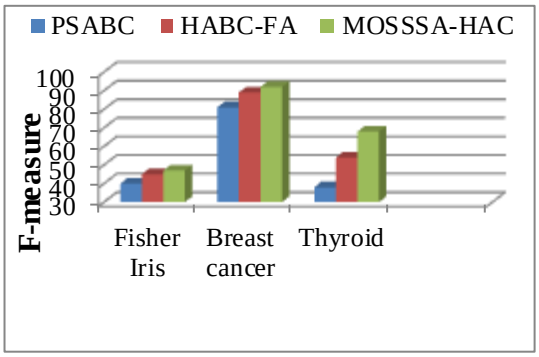

Figure 3 F-Measure

VIII. CONCLUSION AND FUTURE WORK

To achieve an efficient clustering performance, the pre-processing, feature selection and classification process have been performed by using optimized based clustering techniques. In the main research, a hybrid algorithm of PSO and ABC algorithm is utilized for benchmark classification problems. The PSABC algorithm accomplishes nearby too global search performance all the more successfully. Novel clustering algorithm HABC-FA that joins the essential conduct of a FA joined with ABC, which is to enhance the arrangement of clustering problem. The effective and efficient algorithm is proposed is termed as MOSSSA-HAC which is utilized to give more exact clustering results. HAC is connected for high quality clusters based neighborhood likeness esteems. Results demonstrate that the proposed strategy accomplishes high clustering execution regarding exactness, review and f-measure when contrasted with the current technique. In future work the scalable clustering can be connected with hybrid clustering approach way to deal with handle the anomaly information all the more effectively. The future work will investigate the acceleration issue of the ABC-K-Modes.

REFERENCES

[1]. Abraham, Ajith, Swagatam Das, and Amit Konar, 2007. "Kernel based automatic clustering using modified particle swarm optimization algorithm." InProceedings of the 9th annual conference on Genetic and evolutionary computation, pp. 2-9. ACM, 2007.

[2]. Hassanzadeh, Tahereh, and Mohammad Reza Meybodi, 2012. "A new hybrid approach for data clustering using firefly algorithm and K-means." Artificial Intelligence and Signal Processing (AISP), 16th CSI International Symposium on. IEEE, 2012.

[3]. Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, Angela Y Wu, 2002. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE transactions on pattern analysis and machine intelligence 24.7: 881-892.

[4]. Xiaohui Yan, Yunlong Zhu, Wenping Zou, and Liang Wang, 2012. "A new approach for data clustering using hybrid artificial bee colony algorithm." Neurocomputing 97 : 241-250.

[5]. Yunfeng Xu, Ping Fan, and Ling Yuan, 2013. "A simple and efficient artificial bee colony algorithm." Mathematical Problems in Engineering 2013.

[6]. Sangeetha, J., and V. Sinthu Janita Prakash. "An Efficient Inclusive Similarity Based Clustering (ISC) Algorithm for Big Data." Computing and Communication Technologies (WCCCT), 2017 World Congress on. IEEE, 2017.

[7]. Tanır, Deniz, and Fidan Nuriyeva. "An effective method determining the initial cluster centers for K-means for clustering gene expression data." Computer Science and Engineering (UBMK), 2017 International Conference on. IEEE, 2017.

[8]. International Journal of Scientific Research in Computer Sciences and Engineering (ISSN: 2320-7639)

[9]. International Journal of Scientific Research in Network Security and Communication (ISSN: 2321-3256) .

[10]. Gan, G., J. Wu, and Z. Yang, 2009. "A genetic fuzzy k-Modes algorithm for clustering categorical data." Expert Systems with Applications 36.2 : 1615-1620.

Authors Profile

S Karthikeyan pursed Bachelor of Science from Bharathiar University of Coimbatore, in the year 2006 and Master of Science from Bharathiar University in the year 2008. He is received his Ph.D in Computer science from Karpagam University in the year 2019 and currently working as Assistant Professor in Department of Computer Science, Rathinam College of Arts and Science, Affiliated to Bharathiar University,Coimbatore, Tamil Nadu. He has published more than 06 research papers in reputed international journals and conferences and it's also available online(2 Scopus Journal). He is Received an award Nation Builder Award from Rotary Club. His main research work focuses on Data Mining, Clustering, Classification, We based learning, Higher Education Institutional Strategies for Implementing Outcome Base Education.

Mr A. Dhakshina Moorthy pursed Bachelor of Science and Master of Science from Bharathiar University, Coimbatore in year 1896. He also Completed PGDCA. He is currently Freelancer – Data Scientist**.** Overall 25+ years of experience in IT roles spanning Practice Management, Delivery Management, Project Management, Solution Architecting & Client Management. Have good mix of Practice Building, Presales, Delivery and Implementation of BI and Analytics projects. Strong Technical expertise in end to end Analytics implementations using range of Oracle products and Open source products. Having 5 different company working experience, Patni Computers, Pune (2 Yrs), HCL Perot , UK (8 Yrs), Perot Systems , Blr (4 Yrs), Tech Mahindra, Blr (6 Yrs).

