International Conference on Advancements in Computing Technologies - ICACT 2018 ISSN: 2454-4248
Volume: 4 Issue: 2 134 – 138

# Multi-Objective Optimization for Clustering with Evolutionary Systems – A Review

R.Abitha[1], S.Mary Vennila[2]

Research Scholar, Assistant Professor,Department of Computer Science,Women's Christian College, Chennai,India[1]

Associate Professor, PG &Research Department of Computer Science,Presidency College, Chennai, India[2]

**Abstract**:Clustering is a prevalent unsupervised data mining technique for subdividing the given data set into homogeneous groups based on similarity metrics. The traditional clustering algorithms used by many researchers to optimize a single objective function suitable to produce clusters from the given data set tends to converge to local optima which can be resolved by Evolutionary Multi Objective Optimization . EMOO are evolutionary system used to optimize various measures of evolving system. The problem of generating cluster by optimizing specific features is said to be multi objective problem. The objectives to be optimized is connectedness, separation, accuracy and shape etc. The performance of these EMOO is inclined by evolutionary techniques used, objective taken for optimization, chromosome representation, population size reproduction operator used and stopping criteria. These evolutionary policies are widely used by Genetic Algorithm (GA) to generate quality cluster. This paper aims to provide the necessity of EMOO and may pave a path to design a framework for multi objective optimization problem using Evolutionary strategies for clustering in data mining process.

**Keywords**:*Clusters, Evolutionary Systems, Multi Objective Optimization, Genetic Algorithm(GA).*

*****

## I. INTRODUCTION

Data explosion poses challenging research options for useful information retrieval to enhance decision making. There are zeta byte of data transaction available every day. Researchers are discovering new techniques for retrieving interesting patterns and knowledge hidden in the huge amount of data available around us. Data Mining techniques compared with other techniques have been proved as a knowledge discovery technique which results in prediction of reliable patterns for any problem domains by eliminating redundancy. One among the useful data mining techniques is clustering. Clustering is one of the data mining techniques used for multivariate data analysis. Since clustering is unsupervised and an exploratory data analysis it is most apposite for multi objective optimization problems. Due to considering problem under an unsupervised nature the structural features are not known which needs some domain knowledge should be known in advance.

*Clustering as a multi-objective optimization problem*

*Clustering* is an unsupervised technique that group the data objects having similarities with minimum distance between the data objects in a same cluster by eliminating dissimilar data objects from all the clusters including outlier. In data mining, types of clustering algorithms available are, Hierarchical clustering, Partition clustering, Fuzzy clustering etc. Partition based clustering works well with smaller dataset with single objective optimization. Single objective optimization is a technique aims to find best solutioncorresponding to maximum or minimum value of single objective function and fails to provide trade-off between different objective function. Many real life problems expects to be solved by achieving different objectives. Decomposing the given data set into different clusters and maximize or minimize different objectives in parallel is multi objective optimization of clustering in data mining.

Including only one objective function to measure the fitness of cluster will not reflect correct partitioning. Best partitioning of clustering can be based on different aspects including compactness of the clusters and cluster symmetry, density etc. In many problems objectives under consideration conflict with each other and optimizing a single solution with subject to single objective may result in unacceptable clustering. In data mining to provide best and acceptable result of clustering is possible with MOO technique. In MOO Maximization or minimization can be done depending on the problem taken.

*Evolutionary algorithms for solving multi-objective optimization problems*

Evolutionary computing(EC) is an area of computer science where Darwin's Principle of evolution is used to solve problems. This Evolutionary computing extends wide range of algorithms comprising Evolutionary Algorithms(EA), Genetic Algorithms, Evolutionary Programming(EP), Genetic Programming(GP), Gene expression programming, Cultural Algorithm(CA), Swarm intelligence and so on. In this broad area of EC, Evolutionary Multi-Objective Optimization (EMOO) has become a popular and useful field of researchand application.Multi objective optimization is the one optimizing different number of objectives simultaneously. Since different objectives have different optimal solution these objectives are of conflict with others and can be solved with or without any constraints. For optimizing different objectives in MOO problem of clustering, Evolutionary algorithm is suitable for their effectiveness and robustness in searching a set of trade-off solutions. In EMOO, instead of single solution a set of pareto-optimal solutions exist that simultaneously optimizes each objective. Since the evolutionary algorithm uses population approach in the search procedure and it produce number of solutions in a single run it is the best suitable approach to find the optimal result from the pareto-optimal solution.[1]. As the Evolutionary algorithms are robust search methods and adopt to the environment and discover interesting hidden patterns that will be missed by greedy algorithms[2],EO procedure is a perfect choice for solving multi objective optimization problems [1]

 ISSN: 2454-4248
134 – 138

## II. EVOLUTIONARY MULTI-OBJECTIVE CLUSTERING SYSTEM-AN OVERVIEW

### A. Basics of Clustering

Clustering is unsupervised technique in data mining for dividing the population of entities into a number of groups such that objects in the same group are more similar to each other. The aim is to set apart groups with similar traits and assign them into clusters. Clustering can be divided into two groups namely *Hard Clustering* where each object either belongs to a cluster completely or not, *Soft Clustering*where the probability of the data point to be in those clusters is assigned instead assign each entity in separate clusters.[27].Clustering is widely used in many application domains including biology, marketing, economics, medicine, anthropology etc. The need for clustering is to formulate algorithm to deal with large data base, to handle numerical, categorical and binary data, to handle noisy data for finding outlier, finally result produced should be interpretable, and usable.

### B. Evolutionary Algorithms for Clustering

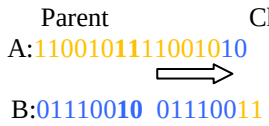
Though there are varieties of evolutionary methods applied for solving different problems in data mining, recurrently used strategies in clustering are Genetic Algorithm(GA) A survey was done on Evolutionary algorithms for clustering by A. Freitas [2]. Survey gives brief idea on as clustering is unsupervised in nature it is important to consider multiple objectives when evaluating the fitness of an individual representing candidate in clustering solution.GA comes under evolutionary algorithm are highly capable of performing successful clustering and good in producing proper number of clusters during the process with small to different large data sets.[3]The clustering technique is influenced by various factors including the initialization of population, chromosome representation, operators used for reproduction including selection, cross over, mutation, selection of population for next generation , criteria for optimization fitness function, and finally area of application.

### C. EMOO System with GA

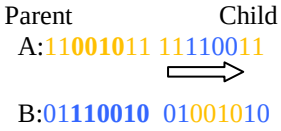
*Chromosome representation:* In GAs, chromosomes are encoded as binary strings. Some systems have fixed length chromosomes and some have variable length chromosome. Acollection of strings called population. Initially a random population is created. As GAs, evolve over generations in each generation they produce new population by applying the genetic operators namely selection, mutation and cross over. For a given problem there are possible set of solutions available. Each solution in the search space is associated with a fitness value based on the value of objective function to be optimized.

*Selection*: Selection process is needed to produce optimal solution by selecting two chromosomes with highest fitness value. Different types of selection mechanisms are available to select the chromosomes in the search space .In *Roulette wheel selection* each individual is given chance to become parent in proportion to their fitness. In *Rank Selection* Chromosomes are sorted according to their raw fitness and ranked hence based on the ranking parents are selected to reproduce the next generation. In *Tournament selection* a set of chromosomes are randomly chosen and with the fittest chromosome will be allowed for mating process.

*Cross Over*:Cross over is applied on selected individuals (parents) by choosing cross over points. In single point cross over random cross over point is selected and tails of both the chromosomes are swapped and produce new off springs.




In two point cross over two positions are randomly selected and the middle portion of the parents are swapped.




*Mutation:* Mutation makes changes to the gene of the chromosome by changing the value of an attribute if value encoding is used and if bit encoding is used then bit flipping is carried out. The mutation point is selected through heuristics.

### D. Issues in multi-objective evolutionary clustering:

The aim of clustering is mainly grouping of the given data set in to different disjoint group where similarity within the same group will be more and less in between the groups. This clustering technique is though very popular data mining technique and used by many researchers in various fields still the evaluation of cluster quality is a provocative . None of the non-GA based algorithms are capable of efficiently and automatically forming natural groups from all the input patterns especially when the number of clusters included in the data set tends to be large.[4] Most of the clustering techniques are based on single objective which result in a single measure of goodness of separating the clusters. The main goal of multi objective optimization is maximizing or minimizing multiple objectives the user is interested in particular under different constraints in parallel. The aim of our proposed work is to provide a system with large set of data with different set of objectives, automatically identifies optimal cluster which satisfies those objectives. The perfect cluster analysis should be scalable, need minimalinput parameter, robust, result independent of data input order. Since these issues cannot be resolved with normal traditional clustering algorithm which is converge to local optima and not to global optima. Based on impossibility theorem defining a generalized framework for clustering method is impossible[5]. It is necessary to have the domain knowledge to choose the clustering algorithm.[5] Depending on the domain the input parameters like number of clusters ,optimization criterion, termination conditions etc. will differ for various clustering algorithms. As a result different clustering algorithm will produce different result for same data set and same algorithm for different data set also will produce different cluster results as there is no single universally accepted clustering algorithm. Due to all these above said reasons the final output of the clustering techniques in data mining which includes different clusters to be validated for the goodness of the clusters. As the intra cluster variation should be minimized and inter cluster variation be maximized, validation of clusters can be done using two techniques namely

WSS ,Within Sum of Squared Error and BSS, Between Sum of Squared Error to achieve compactness and . separation. To validate clusters in various aspects also different techniques are used , they are 1)External Index evaluates the resultant clusters by using prior knowledge /class labels 2)Internal Index evaluates how well the clusters are appropriate for the data without referencing external information. It uses only the given data.3) Relative Index evaluates the result with different parameter values for the same algorithm , mainly used to find optimal number of clusters. Hence there is no specific clustering algorithm[5] the best known evolutionary techniques can be combined with the standard clustering algorithm to evolve the proper number of clusters and to provide appropriate clustering.[4] .

## III. LITERATURE SURVEY

With reference to Dr. K. Meena et al[8] different clustering techniques and different distant measures are available which can be applied on real data sets. The author has used K-Means clustering algorithm and Manhattan distance to calculate the distance between the objects. This method converge to local optima because the author has concluded that the resulting clusters are different for different seed values. So the final cluster structure entirely relies on the choice of initial seeds. Hence it is forced to determine the system converge to global optima.

Sridevi Radhakrishnan et al.,[9] confirmed that selection of data mining technique depends on the nature of domain with data set. For labeled data set classification technique is suggested for best prediction. For unlabeled data set clustering is best suited for pattern recognition. For optimization of results bio inspirational based technique isbest suited. Evolutionary algorithm based clustering technique can be derived to optimize a problem.

Bala Sundar V et al.,[10] adopted K-Means technique and Euclidean distance in prediction of heart disease diagnosiswith real and artificial data sets. Cluster compactness and connectedness measure are achieved with high speed and accuracy rate more than Decision tree, Naive Baye's and Neural Network

Arpit Bansal et al.,[11] presented in his paper an improved K-Means algorithm for prediction analysis applied for cancer dataset. In this work author has enhanced the Euclidean distance formula by implementing normalization. The proposed modification achieved 92.87% accuracy for large data sets with quality clusters.

The problem of mild cancer level in Erode district was brought in to lime light through the research of Lookman sithic et al.,[12].The real time data set consisting of 424 instances with 10 different attributes were taken for their research They adopted K-Means algorithm with Euclidean distance measures which produced the result in the iteration 4.They have concluded that this method had trained the data up to 100%and so the error rate also reduced relatively with less time.

Bouhmala et al.,[13] proposed a combination of k-Means with genetic algorithm for clustering problem. The idea is to use genetic search approach to produce clusters using two point cross over and then applying k-means technique to improve the quality of cluster in order to speed up the search process. The proposed algorithm converges faster while generating the same quality of clustering compared to genetic algorithm.

Helan Cynthiya Y.,et al.,[15] presented an Enhanced K-Means Genetic algorithm (EKMGA)finds the fittest value of the chromosome for the breast cancer and lung cancer data sets with minimum number of generations.

Anusha M et al.,[16] have done performance analysis on two meta heuristics algorithms Multi objective Genetic algorithm and Particle Swarm Intelligence with real–life data sets clustering problem. They inferred that MOGA is faster than MOPSO in terms of effectiveness and efficiency to find the optimum result.

Anusha M et al.,[17] proposed an improved feature selection algorithm using K-Means Genetic algorithm for multi objective optimization problem. FS-NLMOGA maximizes two objective functions and minimizes an objective function simultaneously. The quality of the cluster is increased with high accuracy than NLMOGA. The author tested the performance of the proposed algorithm with several real life benchmark data sets.

Anusha et al.,[18]proposed evolution clustering multi objective optimization[ECMO] algorithm In this work, the best feature of the dataset was identified using selecting features (CL) of criterion learning algorithm which produce the resultant cluster with high accuracy and predict appropriate number of clusters. This algorithm uses Rand-Index to measure the validity of clusters.

From the above literature survey, different clustering algorithms and the techniques involved are presented in Table I

TABLE I DIFFERENT TYPES OF CLUSTERING ALGORITHMS

| Algorithm / Technique used | Reference Paper | Test Data Used | Purpose of clustering | Advantages | Limitations / Future Research |
|---|---|---|---|---|---|
| K-Means method uses Manhattan distance measure | Dr. K. Meena et al.,IJISA- [2013] | Real time student data set | Clustering with real data and to find outlier | Presented different clustering techniques and distance measures.<br>Outlier also detected | Converge to local optima. Resulting cluster depends on the selection of seed value.<br>Need a system which converge to global optimal result. |
| Study on data mining techniques in health care | Sridevi Radhakrishnan et al.,IRJET-[Aug | Health Care dataset | To analyze data mining techniques in | Concluded with suitable techniques for different labeled and unlabeled data sets. | Need to derive a system for clustering in unsupervised nature using bio inspired techniques. |

| | | | | | |
|---|---|---|---|---|---|
| | 2015] | | diagnosing heart disease, breast cancer, diabetes and Liver disease | | |
| K-means clustering technique | Bala Sundar V et al., IJCA[2012] | Real and Artificial dataset | Prediction of heart disease diagnosis | Accuracy rate is more than Decision tree, Naive Baye's and Neural Network within less time.<br>Ensures compactness and connectedness of a clusters | Combination of K-means and MOO can be used to improve the result. |
| K-Means with Normalization technique | Arpit Bansal et al.,-IJCA[2017] | Cancer Dataset | Prediction Analysis | Accuracy is Improved and clustering time is reduced | Applicable for smaller data sets. Suitable System isneeded for large data sets. |
| K-Means with Euclidean distance measures | Lookman sithic et al.,IJIES [2015] | Real data set with 424 instances | Identify affected people's cancer level | Result produced in iteration 4. Error rate reduced with less amount of time | Not mentioned about cluster validation Which is needed very much as the system used K-Means .<br>More accurate results can be produced with Evolutionary Systems |
| K-Means with Genetic Algorithm | Bouhmala et al.,-IJMO[April2015] | UCI data sets Iris, glass, Seeds ,wine , Habermann, Fertility, Blood | To improve cluster quality and speed up search process | Faster convergence and produced quality cluster as GA | Search process still can be improved with evolutionary system as they search parallel. |
| Evolutionary K-means Genetic Algorithm-EKMGA | HelanCynthiya Y.,et al., IJARCET-[2015] | Cancer data set from UCI | Clustering of cancer genes | Provides fittest value for the datasets with minimum number of generations | To provide better optimal solution multi objective genetic algorithm (MOGA) can be used which works on parallelism mechanism. |
| MOGA and MOPSO | Anusha M et al., IJMCS-[2015] | Real-life data sets from UCI Machine Learning Repository | Performance Analysis | MOGA is effective and efficient to find optimal result than MOPSO | Can develop a hybrid method to achieve high effectiveness, efficiency, and consistency simultaneously |
| Improved feature selection algorithm using k-means genetic algorithm FS-NLMOGA | Anusha M et al., Elsevier[2015] | Real life data sets from UCI Machine Learning Repository | To maximize compactness and accuracy of clusters through constraint feature selection | Simultaneously optimize the chosen objectives with high accuracy | to improve the time complexity system can be enhanced with constrained crossover on high dimensional data sets |
| Evolution clustering multi objective optimization with CL algorithm | Anusha M et al., IJACSA[2016] | Prima Indian diabetes data set from UCI | Cluster with High accuracy | Best feature of the data set identified and predicted number of clusters | System can be improved by optimizing more number of objectives. |

## IV. SUMMARY AND FUTURE RESEARCH CHALLENGES

A. *Summary:*Data mining is a process of finding interesting hidden pattern from the available data set. Cluster analysis is one of the data mining unsupervised technique in which objects in the same cluster have similar features. The goal of clustering is to provide natural high quality clusters with high similarity measures within cluster and high dissimilarities between the clusters. Non- GA clustering algorithms like K-Means partition algorithm are inefficient in finding natural groups from the input pattern[3] especially when the dataset is large and also suffer from local optima though it is one of the well-known simplest partition clustering algorithm . Many real life problems are to be solved by achieving different objectives. The best choice is Evolutionary based algorithm can be combined with the traditional partition algorithm which finds globally optimal solution for multi objective optimization problems since evolutionary based algorithms are stochastic search technique that performs multi directional search. Evolutionary algorithms particularly GA based algorithms are stimulated by "survival of fittest"[Darwin's theory of natural evolution]. That is from the initial population through several evolutionary steps a set of new more appropriate solution are attained that leads to near optimal solution. They perform intelligent search so these evolutionary based GA can be used in determining the initial value of the cluster centroid and leads to better result than traditional partition K-means algorithm where the initial cluster centroid chosen randomly.

B. *Future Research Challenges:*Many researchers have done their research in cluster analysis using different clustering algorithms in various domain. But the research based on traditional partition clustering algorithms converge to local optima and provide best clusters based on optimizing single objective. Since there is no universal algorithm for clustering[5] the clustering algorithm and parameters are to be chosen based on the domain. Due to these reasons recent decades many researchers are showing interest in Multi objective optimization problem. For these MOO problems more than one objectives to be optimized simultaneously, these traditional algorithms are not suitable hence researchers are implementing evolutionary based algorithm to solve MOO problems. According to the review

researchers have not discussed about the memory for any of their work. Since Evolutionary algorithms are termed as memory less as they do not retain memory of previous generations. From these review the research directions identified are i) what happened to the elite individuals of each generation. ii) Is there any specific algorithm to dealt with and to store these elite individuals iii) The framework by combining Evolutionary algorithm with traditional clustering algorithm to be designed to optimize many objectives in parallel and generate best optimal quality cluster along with memory storage for elite individuals .

## REFERENCES

[1] Kalyanmoy Deb"Multi-objective Optimization Using Evolutionary Algorithms: An Introduction" In Wang L., Ng A., Deb K.(eds) Multi-objective Evolutionary Optimization for Product Design and Manufacturing. Springer,London,pp.3-34,2011

[2] Eduardo Raul Hruschk, Ricardo J.G.B, Alex A. Freitas,André C. P. L. F. de Carvalho, "A Survey of Evolutionary Algorithm for Clustering",IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews ,Vol. 39,March 2009.

[3] MamtaMor Poonam Gupta , "A Review on Clustering with Genetic Algorithms" International Journal of Computer Science and Communication Networks,Vol.4(3),pp.94-98,June 2014.

[4] Rahila H.Sheikh,Raghuwanshi M.M,Anil N.Jaiswal,"Genetic Algorithm Based Clustering:ASurvey",IEEE-First International Conference on Emerging Trends in Engineering and Technology,pp.314-319,2008.

[5] Jyothi,Neha Kaushik,Rekha,, "Review Paper on Clustering and Validation Techniques", International Journal For Research in Applied Science and Engineering Technology(IJRASET),Vol.2,IssV,May 2014.

[6] L. L. Kaufman and P. J. Rousseuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.

[7] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, New York, 1989.

[8] K. Meena, , M. Manimekalai , and S.Rethinavalli , "An Unsupervised Technique for Statistical Data Analysis Using Data Mining" International Journal of Information Sciences and Application,Vol.5, pp.11-20,2013

[9] Sridevi Radhakrishnan, Shanmuga Priyaa S, "A Critical Study of Data Mining Techniques in Health Care Data Set", International Research Journal of Engineering and Technology,Vol. 02,August 2015

[10] Bala Sundar V, Devi T, Saravanan N, **"**Development of a Data Clustering Algorithm for Predicting Heart Disease",International Journal of Computer Applications, Vol .48,June 2012.

[11] Arpit Bansal, Mayur Sharma, Shalini Goel,"Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining", International Journal of Computer Applications,Vol.157, January 2017.

[12] Lookman Sithic H, Uma Rani R**,**"A Grouping of Cancer in Human Health using Clustering Data Mining Technique", International Journal of Inventive Engineering and Sciences,Vol.3,May 2015.

[13] N. Bouhmala, A. Viken, and J. B. Lønnum,**"** Enhanced Genetic Algorithm with K-Means for the Clustering Problem"International Journal of Modeling and Optimization,Vol.5,April 2015.

[14] Bini B.S,Tessy Mathew, "Clustering And Regression Techniques For Stock Prediction", International Conference on Emerging Trends in Engineering, Science and Technology(ICETEST) –Science Direct, Procedia Technology,pp.1248-1255,2015.

[15] Helan Cynthiya Y., Anusha M., Dr. J. G. R. Sathiaseelan, " An Unsupervised Classification of Cancer Genes Using Genetic Algorithm", International Journal of Advanced Research in Computer Engineering& Technology,Vol.4,Iss 7,July 2015.

[16] M.Anusha and J.G.R.Sathiaseelan, "Cluster Performance Analysis of Multi-objective Genetic Algorithm and Particle Swarm Optimization Techniques", International Journal of Modern Computer Science,Vol.3,Iss 2,June 2015.

[17] M.Anusha and J.G.R.Sathiaseelan, "Feature Selection using K-Means Genetic Algorithm for Multi-objective Optimization",ScienceDirect- Procedia Computer Science 57 ( 2015 ), pp.1074 – 1080.

[18] M.Anusha and J.G.R.Sathiaseelan, "Multi-Objective Optimization Algorithm to the Analyses of Diabetes Disease Diagnosis", International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 7, No. 1, 2016

[19] Bhuvaneswari K, Anusha M, Sathiaseelann J.G.R, "A Comparative Analysis of Clustering Techniques using Genetic Algorithm", International Journal of Computer Science and Mobile Computing,Vol .4,Iss 5,pp.80-86, May-2015.

[20] M.Anusha and J.G.R.Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", International Journal of Applied Engineering Research, pp. 228-231, 2015.

[21] M.Anusha and J.G.R.Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", International Journal of Applied Engineering Research, pp. 228-231, 2015.

[22] Ke Li, Sam Kwong, Kalyanmoy Deb, "A dual-population paradigm for evolutionary multi objective optimization.", Information Sciences, Elsevier, 2015.

[23] Sabhia Firdaus, Ashraf Uddin Md,"A Survey on Clustering Algorithms and Complexity Analysis", International Journal of Computer Science Issues, Vol. 12,March 2015.

[24] Rachsuda Jiamthapthaksin,Christoph F. Eick,Ricardo Vilalta,"A Framework for Multi-objective Clustering and its and Co-locationMining," Advanced Data Mining and Applications. ADMA 2009,Vol 5678. Springer, Berlin, Heidelberg, pp.188-199,doi.org/10.1007/978-3-642-03348-3_20

[25] C. A. Coello Coello, G. B. Lamont, and D. A. van Veldhuizen, *Evolutionary*Algorithms for Solving Multi-Objective Problems, 2nd ed., ser.Genetic and Evolutionary Computation. Berlin–Heidelberg, Germany:Springer, 2007.

[26] Hartono, Erianto Ongko, Dahlan Abdullah "Determining a Cluster Centroid K-Means Clustering Using Genetic Algorithm", International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 6, pp.160-164,June 2015

[27] An Inroduction to Clustering-www.analyticsvidhya.com.