Data Clustering for Optimized Information Search with Hybrid Evolutionary Approaches

Anuradha D. Thakare

Abstract: ***Clustering is an important data analysis technique which reveals the relationships among unexplored data objects. Cluster initialization and selection of seeds in first iteration contributes to the quality of clustering. The prime objective is to find best cluster with some quality measure. K-means is prone to local optima since initial centroids are selected randomly. In order to evaluate this problem, some heuristic clustering algorithms are introduced along with evolutionary approaches like Genetic Algorithms and Swarm Intelligence. Genetic Algorithms are the heuristic search techniques and are found to be robust to envisage the optimal or near optimal combination of weights in a multidimensional space. This article presents comparative analysis of various hybrid evolutionary approaches developed for clustering to find the optimal cluster center. The objective is to improve the quality of clusters. From the analytical and experimental results, it is observed that the proposed hybrid evolutionary algorithms perform satisfactorily over the existing approaches. As compared to hybrid PSOBA, Multi Stage Genetic Clustering results into reduced error rate by 30 to 50 percent for thyroid and iris dataset respectively. The clustering results vary with respect to dataset and the internal spread.***

Keywords: clustering; evolutionary algorithms; Genetic Algorithms(GA); Particle Swarm Optimization(PSO); Bee Algorithm(BA); K- means(KM).

I. INTRODUCTION

The clustering algorithms are sensitive to the choice of a cluster centers. K-means is a partitioned clustering algorithm, which clus-ters the objects and represents it by the mean value of the resulting partition. The iterative nature of k means algorithm scans the entire input and objects multiple times, and falls into local optimum giving the results mostly accumulated to the selection of the initial seeds. Several methods have been proposed to overcome these limitations of K-means. However, none of these methods examines the intra-cluster and inter-cluster distances simultaneously to overcome the problem of cluster initialization. The global convergence ability of Genetic Algorithm overcomes the limitation of K-means.

Genetic algorithms are good in searching a multidimensional space and finding the optimal solutions. This motivated the use of Genetic Algorithm for the proposed research to search for an optimal or near optimal combination of weights.

The aim of this research work is to suggest the evolutionary ap-proach for clustering to find the optimal (best) cluster centers whereas the first objective was to identify the clustering problem as an optimization task. In order to get optimal clusters, traditional clustering algorithms failed especially when search space is com-plex. Therefore another objective of this research was to design a computational model with heuristic search approach and develop the hybrid algorithms utilizing global search capability of the algorithm. The powerful and domain independent search ability of GA is utilized for the said research objectives. The research solution provided is a simple and easy to use system to handle the high dimensional data using the evolutionary algorithm, which facilitate swift and pertinent text retrieval without any loss of vital in-formation.

II. EVALUATION OF RELATED RESEARCH

Many of the researchers have solved the clustering problem by using heuristic approach. Different evolutionary approaches are reported in the literature. For the data clustering, many hybrid algorithms are provided in the research using evolutionary algo-rithms like Genetic Algorithm and Swarm Intelligence. In our work, we have investigated the same and suggested the optimization algorithms which give better results in certain situations. This research work aims towards producing the optimized clusters thereby improving the quality of clusters. The approach applied in the proposed research work is divided at two levels; first we de-veloped the new data clustering methods using evolutionary tech-niques i.e. GA and PSO which performed significantly better than existing methods for data clustering and secondly, we suggested hybrid information retrieval model using GA for efficient retrieval. Also new matching functions are developed for information re-trieval which performs better over existing matching functions. The performance is compared with existing methods over benchmarked datasets. After review, following challenges in Clustering are identified:

Parameters: Clustering algorithms require certain parameters for clustering such as the no. of clusters, cluster shapes, orientation of cluster etc.

Initialization: Due to a bad choice of initial cluster centers often clustering algorithms converges to local optimum. This results in the poor quality clusters.

Dimensions: Reducing the dimension of clustering space

III. OVERVIEW OF HYBRID EVOLUTIONARY AP-PROACHES FOR DATA CLUSTERING

In this research work, seven evolutionary approaches are proposed. four for clustering and three for Information Retrieval. These approaches are hybrid models which are developed with a new methodology and considering best contribution from primary algorithms.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

All contributions achived the results in terms of accurate Clustering and efficient Information Retrieval. Table 1 depicts the brief overview of Hybrid Clustering approaches.

Table 1: Overview of Hybrid Clustering Approaches

Approach	Strength	Remarks	Existing Approach
Improved K-means Clustering (IKM), Elsevier, 2014	Intracluster And intercluster distance are simultaneously optimized.	The entire input dataset is scanned only once to assign the cluster membership	Improved k-means hybrid model is proposed
Multi_Stage Genetic Clustering (MSGC) IEEE, 2015	GA is designed with multiple objective functions for clustering	Two stage GA for accurate clustering	AMOSA is used for initial partitions
Two Stage Genetic KHM (TSGKHM) Clustering Springer, 2014	KHM is used as a fitness function in Genetic Algorithm.	Two stage GA to avoid local optima and results in optimal clusters	KHM is integrated with only SI approaches
Swarm Intelligence based hybrid method for data clustering IEEE, 2013	Method performs a local and global search simultaneously .	Integrates the behavior of PSO, and BA to produce accurate clusters	Hybrid methods based on PSO, ACO, GA and K-means

The researchers have proposed different approaches with some assumption and constraints to solve clustering and in-formation retrieval problems. This research work investigated and developed some approaches in line with earlier research that are summarized in following paragraphs.

3.1An improved K-means[IKM] based method for data clustering

An improved K-means based method for data clustering is pro-posed [1] which can automatically partition the entire data points and produces the optimal clusters. The proposed method minimizes an intra-cluster scatter and at the same time maximizes an inter-cluster scatter. All the objects are evaluated based on simultaneous execution of the two objective functions, minimum intra-cluster scatter, and maximum inter-cluster scatter. This assigns the objects exactly once to the cluster and optimizes the number of cycles required to reach the exact partitions. Two objective functions produce pairs of values such that in a cluster, the distance between points is minimium, and the inter-cluster distance is maximum. The performance is compared with three clustering algorithms. It is observed that the proposed algorithm produces more accurate results and is well suited to almost all the data sets. Error rate is reduced giving more accurate clusters, and the improvement is seen in F-measure values.

3.2 Multi Stage Genetic Algorithm for Data Clustering [MSGC]

A new Multi Stage Genetic Algorithm for data clustering is pro-posed [2], which can automatically partition the entire data and produces the optimal clusters. Each partition is divided into several small sub clusters for the optimization purpose. The objective functions are evaluated for merging the sub clusters [3] and produce the final accurate clusters. The performance of the proposed MSGC is compared with three clustering algorithms. The result of the proposed algorithm is more accurate and well suited to almost all the data sets except the liver disorder and the sph_4_3 dataset. It is seen that the Intracluster distance reduces drastically giving more accurate clusters, and the average error rate is also reduced by eight to ten percent. A lot of work is needed to add more objectives and get good clustering by using a parallel genetic algorithm.

3.3. Hybrid Swarm Intelligence Method for Data Clus-tering[PSOBA]

A new hybrid Swarm Intelligence method for data clustering using PSO and Bee algorithms is introduced [4]. This hybrid clustering algorithm performs outstanding in terms of accuracy in clustering results with minimum intra-cluster distance. Experimental results demonstrate that the proposed method does not trap at local optimal space as it uses Bee Algorithm for clustering. The minimum intra-cluster distance is used as a metric to search the robustness of data cluster centers in N-dimensional Euclidean space [5,6]. We got results of the hybrid algorithm better than K-Means in terms of error rate since PSO and BA performs well to solve the problem of local optima. For comparison purpose, same numbers of evalua-tions are conducted on other algorithms like simple PSO [7], BEE, and K-means algorithms. Each time new solution is executed by the fitness function.

IV. COMPARISON OF HYBRID EVOLUTIONARY AP-PROACHES FOR CLUSTERING, MSGC AND PSO-BA

This section discusses the comparison of two hybrid evolutionary approaches MSGC and PSOBA for data clustering whereas; more focus is on percentage improvement.

The experimentations are carried out and results are compared with existing approaches on standard datasets. The methodology adopted for every clustering method differs and are compared using various performance measures. The evolutionary approaches for clustering are MSGC and PSOBA. These are compared based on percentage error rate as depicted in Figure1. These clustering methods cannot be compared as per Intracluster distance as MSGC work on multiple clustering objectives to produce compact clusters.

The distance reduced drastically as compared to PSOBA almost for all the datasets. The comparison of all contributions for clustering in terms of percentage error rate is given below:

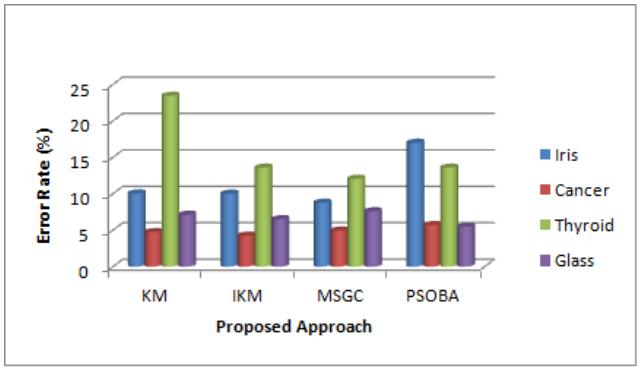

Fig. 1. Performance comparison of hybrid evolutionary algorithms MSGC and PSOBA

V. COMPARATIVE ANALYSIS OF HYBRID MULTI-STAGE GENETIC CLUSTERING WITH OTHER CLUS-TERING METHODS

The clusters present in the real life data sets from UCI Machine Repository [8] are with high dimensions and other datasets with hyper spherical and overlapped clusters [9,10,11,12]. During ex-perimentation, the best values of Intracluster Distance and Error Rate as performance criteria are obtained by four algorithms for all the datasets. MSGC results into accurate clusters among all the clustering algorithms [2, 13, 14] for all the datasets. The percentage improvement of MSGC over other evolutionary algorithms like PSO and BA are reported in Table 2. The minimum values of both the measures reflect good clustering with minimum outliers. Therefore negative values reflect percentage improvement.

Table 2: Results of Multi-Stage Genetic Clustering [2] with KM (K-Means), PSO (Particle Swarm Optimization), BA (BEE Algorithm). Here, Dist is the Intracluster Distance and Error denotes Error Rate

Datasets	Criteria	PSO	BA	MSGC	% Improvement over PSO	% Improvement over BA
Iris	Dist	39.35	33.16	**23.32**	**-40.74**	**-29.67**
	Error (%)	16.44	10.05	**8.78**	**-46.59**	**-12.64**
Wine	Dist	401.01	327.46	**70.01**	**-82.54**	**-78.62**
	Error (%)	6.43	4.17	**3.4**	**-47.12**	**-18.47**
Cancer	Dist	307.27	299.23	**134.75**	**-56.15**	**-54.97**
	Error (%)	5.37	4.42	**4.96**	**-7.64**	**12.22**
Thyroid	Dist	7191.3	7081.7	**721.61**	**-89.97**	**-89.81**
	Error (%)	22.56	20.05	**12.06**	**-46.54**	**-39.85**
Liver Disorder	Dist	8180.9	8016.2	**9639.2**	**17.83**	**20.25**
	Error (%)	8.36	18.95	**22.285**	**166.57**	**17.6**
Glass	Dist	12800.68	12661.48	**5359.21**	**-58.13**	**-57.67**
	Error (%)	6.68	8.62	**7.60**	**13.77**	**-11.83**
Lung Cancer	Dist	13552.93	13552.93	**13552.93**	**0**	**0**
	Error (%)	1.51	1.51	**1.51**	**0**	**0**
Sph_4_3	Dist	246.82	455.99	**49.31**	**-80.02**	**-89.19**
	Error (%)	12.25	9.11	**22.31**	**82.12**	**114.9**
Sph_5_2	Dist	117.72	116	**12.39**	**-89.48**	**-89.32**
	Error (%)	30.27	30.13	**18.16**	**-40.01**	**-39.73**
Sph_6_2	Dist	332.87	332.87	**152.83**	**-54.09**	**-54.09**
	Error (%)	29.11	31.28	**16.67**	**-42.73**	**-46.71**
Sph_9_2	Dist	10.43	10.44	**6.46**	**-38.06**	**-38.12**
	Error (%)	25.25	20.44	**12.62**	**-50.02**	**-38.26**

The proposed MSGC is able automatically to form the appropriate number of clusters. The intracluster distance is reduced drastically, producing accurate clusters for the real life datasets i.e. iris, wine, cancer, thyroid, liver disorder, glass, cancer than other data sets. The results of MSGC shows minimum error rate for almost all the datasets except for, Liver Disorder and Sph_4_3. The percentage improvement in terms of Intracluster distance and error rate is observed almost for all the datasets except Liver Disorder. The negative values for both the performance paramerter reflects that the Intracluster distance is not reduced due to misplacement of some data vectors which inturn results in increased error rate.

5.1 Reasons for improvements in the results

i. Four objective functions for clustering worked for producing the compact clusters. The connected clusters are automatically merged and outliers are removed which results in cluster compactness.

ii. The Sph_4_3 dataset contains highly well separated clusters which are hyper-spherical in shape. A point on the edge of a cluster is closer to some objects in another cluster.

Hence the improvement in error rate is not observed.

iii. In Liver disorder dataset, the first 5 variables are all

blood tests which are thought to be sensitive to liver disorders. The attributes in liver dataset are of type Integer and values are spread from 0 to 200. Hence this dataset has not produced promising results for clustering.

iv. No improvement in the results for Lung cancer dataset is ob-served. This dataset contains 3 classes, 32 objects and 56 features which are almost double the number of objects. Hence algorithm converge early giving same result for all algorithms.

VI. CONCLUSION

This article aimed to address the performance comparison of hybid evolutionary approaches for clustering using evolutionary approaches such as Genetic Algorithm and Swarm Intelligence. Automatic partitioning of the dataset and accuracy of results are the performance criterias. Since it is hard to define accurate clustering or good clustering, every newly proposed solution in the literature is inline with and supersits the earlier work done. In the hybrid evolutionary approach, the best features of well-known heuristic methods such as Genetic Algorithm and Swarm Intelligence are utilized to increase the performance of Clustering. Comparative analysis of two hybrid evolutionary approaches MSGC and PSOBA are discussed and it is observed that MSGC supersits the PSOBA. The Error Rate for MSGC is reduced by 30 to 50 percent for thyroid and iris dataset respectively as compared to hybrid PSOBA. The clustering results may vary with respect to the internal spread of dataset. MSGC is also compared with PSO and BA. The results shows minimum error rate for all the datasets except for, Liver Disorder and Sph_4_3. The percentage improvement of eight to ten percent is observed for all the datasets except Liver Disorder.

REFERENCES

1. A. D. Thakare, C.A. Dhote, An Improved k-means Algorithm with simultaneous optimization of clustering objectives, International Conference on Emerging Research in Computing, Information, Communication and Applications' - ERCICA-2014. Publication in Elsevier and Elsevier digital library, 2014.
2. A. D. Thakare, C.A. Dhote, Novel Multi Stage Genetic Clustering method for multi-objective optimization in Data Clustering, ICCUBEA 2015, Scopus Indexed, IEEE Xplore
3. A. D. Thakare, C.A. Dhote, A Two-Stage Genetic K-harmonic means method for data clustering, Third International Symposium on Intelligent Informatics (ISI' 2014), Advances in Intelligent and Soft Computing (Springer) Series. Volume Title: Advances in Intelligent Informatics.
4. A. D. Thakare, S. M. Chaudhari, Introducing a Hybrid Swarm Intelligence Based Technique for Document Clustering, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 6, November- December 2012, pp.1455-1459 1455
5. C.A. Dhote, A. D. Thakare, S. M. Chaudhari, Data Clustering Using Particle Swarm Optimization and Bee Algorithm, Internaional Conference on Computing Communication and Networking Technologies,2013, IEEE, DOI: 10.119/ICCCNT.2013.6726828 , Page(s): 1-5
6. A.D. Thakare, Dr. C.A. Dhote, S. M. Chaudhari, Intelligent Hybrid Approach for Data Clustering, Advances in Recent Technology in Computing- 2013, IEEE
7. Yanping Lu, Shengrui Wang, Shaozi Li, change, Particle Swarm optimizer for variable weighing in clustering high-Dimensional data, Zhou January 2011, Machine Learning.
8. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
9. S. Bandyopadhyay and U. Maulik, ``Nonparametric genetic clustering: Comparison validity indices", IEEE Transactions on Systems, Man and Cybernetics, Part C, vol. 31, no. 1, pp. 120-125, 2001
10. S. Bandyopadhyay and U. Maulik, ``Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification", Pattern Recognition, vol.35, pp. 1197-1208, 2002
11. S. Bandyopadhyay and S. K. Pal, ``Classification and Learning Using Genetic Algorithms: Applications in Bioinformatics and Web Intelligence", Springer, Heidelberg, 2007
12. S. Bandyopadhyay, C. A. Murthy and S. K. Pal, ``Pattern Classification Using Genetic Algorithms", Pattern Recognition Letters, vol. 16, pp. 801-808, August 1995
13. Xiaoyan CAI, Wenjie Li " A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously", Information Sciences 181 (2011) 3816–3827, ElsevierMenendez H.D.; Barrero D.F.; Camacho, D., A Multi-Objective Genetic Graph- Based Clustering algorithm with memory optimization, IEEE Congress on Evolutionary Computation (CEC), 2013.
14. Clustering data set to categorical feature using a multi-objective genetic algorithm, Dutta, D.; University Institute of Technology., Golapbug, India; Dutta, P.; Sil, J., International Conference on Data Science& Engineering (ICDSE), 2012

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication