ISSN:2249-5789

A Review on Clustering with Genetic Algorithms

MamtaMor Poonam Gupta
OITM, Dept. of CSE, GJUS&TOITM, Dept. of CSE, GJUS&T
mamtamor12121990@gmail.com poonamjindal3@gmail.com

Abstract

This paper presents a review on genetic algorithms based clustering techniques. Clustering is one of the most important tasks of data mining for exploring data sets. It can be used to extract useful and hidden information from the datasets. Clustering techniques have a large area of applications including bioinformatics, web use data analysis and image analysis etc. Traditional clustering algorithms applied to datasets most of the times result in sub-optimal solution due to large search space, so evolutionary algorithms particularly genetic algorithms are best suited for the clustering tasks. The capability of Genetic algorithms is applied to find optimally disjoint partitions and proper number of clusters for a dataset.

1. Introduction

Data mining is theprocess of extracting useful and hidden information or knowledge from data sets. The information so extracted can be used to improve the decision making capabilities of a company or an organization[1][2][3].Data mining consists of six basic types of tasks which are Anomaly detection, Association rule learning, Clustering, Classification, Regression and Summarization. Clustering is one of the important tasks of data mining. Clustering is the unsupervised classification of data objects into groups or clusters. Clustering is defined as the task of grouping objects in such a way that the objects in the same group/cluster share some similar properties/traits.Many non-GA-based algorithms such as K-means and Fuzzy-c-means havebeen used for the clustering tasks [7]. One of the main goals of clustering algorithms is to find 'natural' groups in the dataset along with partitioning the data into those natural groups [2][4]. But none of these non-GA-clustering algorithms are efficient enough to discover 'natural' groups from all the input patterns, especially when the number of clustersincluded in the data set tends to be large. These algorithms also suffer from the problem of local convergence due to large clustering search space [1],[7].

Evolutionary algorithms particularly genetic algorithms have mostly been used as the appropriate algorithms for the clustering task. The genetic algorithms find optimal solutions and globally optimal disjoint partitions of a dataset.A good GA explores the search space properly as well as exploits the better solutions to find the globally optimal solution [5].GAs arethe directed and stochastic search method belonging to the class of probabilistic algorithms. A GAworks on a population of individuals (chromosomes) and produces new population with every generation by applying genetic operators.

This paper is organized as follows: Section 2 and section 3 give a brief introduction on clustering and genetic algorithms respectively. Section 4 contains overview ofexisting GA based clustering techniques and their applications. Section 5 discusses conclusion. Section 6 gives the references.

2. Clustering

Clustering is the task of partitioning the data being mined into several clusters of data objects, in such a way that:

(a) theobjects in a cluster resemble to each other to a great extent; and

(b) the objects of a cluster are much different from the objects in another cluster.

A good clustering algorithm always maximizes the intra-cluster similarity and minimizes the inter-cluster similarity [2],[3],[4]. Among the several types of clustering algorithms, the two most popular are:

i. Hierarchical Clustering methods- produce a hierarchy of clusters
ii. Iterative-partitioning methods- produce a flat clustering solution

Each of these two types can be divided into two subtypes as shown in Fig. 1:

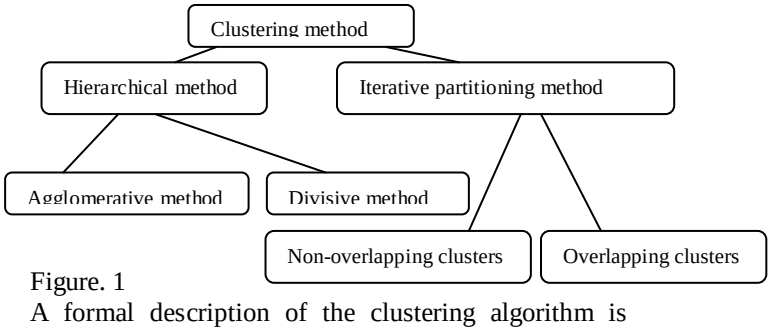

Figure. 1

A formal description of the clustering algorithm is given below:

Let us consider a dataset X= {x_1,x_2, ...x_n} be a set of n objects, each with p attributes. The aim of clustering algorithm is to find K C_1, C_2,…C_k, clusters in such a way that:
(a) $C_i \neq \varphi$for i = 1, . . .k,
(b) $C_i \cap C_j = \varphi$ for i, j = 1, . . .k; $i \neq j$and
(c) $\bigcup_{i=1}^{k} C_i = X$
This means each cluster must contain at least one object, no two clusters can have the same element and all the clusters union should result in X.

Clustering is often based on some similarity or distance measure. The notion of similarity is always problem-dependent.The dissimilarity(or similarity) between the objects is typicallycomputed based on the distance between each pair of objects. These measures include the *Euclidean*, *Manhattan*,and *Minkowski distances*[1].The most popular distancemeasure is Euclidean distance, which is defined as:

$$\mathbf{D(i,j)=}\sqrt{(X_{i1}-X_{j1})+(X_{i2}-X_{j2})+\cdots+(X_{in}-X_{jn})}$$

where i=(x_{i1}, x_{i2}, : : : , x_{in}) and j =(x_{j1}, x_{j2}, : : : , x_{jn}) are two n-dimensional data objects.
Mostly, the objects are clustered on the basis of Euclidean distance. The objects are clustered in such a way that each object belongs to the cluster whose centroidto object Euclidean distance is minimum.
Clustering can be formally considered as a particular kind of NP-hard groupingas far as optimization perspective is concerned. Evolutionary algorithms particularly genetic algorithm are believed to be effective on NP-hard problems. GAs are able to provide near-optimal solutions to such problems in reasonable time. Under this assumption, a large number of genetic algorithms for solving clustering problems have been proposed in the literature. These algorithms are based on the optimization of some objective function (i.e., the so-called fitness function) that guides the evolutionary search.

2. Genetic Algorithms

A GA is a stochastic search technique which performs a multi- directional search by maintaining a population of potential solutions and encourages information formation and exchange between these directions. It takes as input a population of individuals (binary or real valued) which evolves over generation by applying genetic operators (crossover and mutation)[5],[7] as shown in Fig. 2.

A GA for a particular problem must have the following components:
a) a genetic representation
b) a way to create initial population
c) a fitness function
d) genetic operators

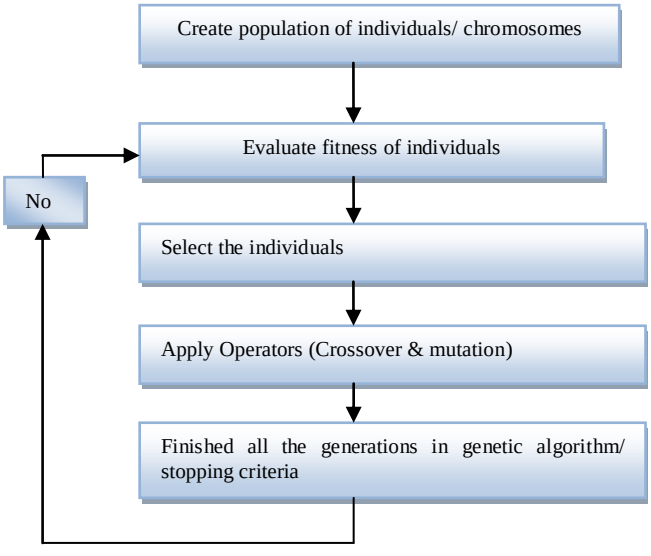

Figure.2

Genetic algorithms are main paradigm of evolutionary computing. GAs are inspired by Darwin's Theory of Evolution-"**Survival of the Fittest**". The general scheme of evolutionary in genetic along with Pseudo-code is shown below:

Pseudo-Code
BEGIN
INITIALIZE population with random candidate solution
EVALUATE each candidate
REPEAT UNTIL (termination condition) is satisfied DO
1. SELECT parents;
2. RECOMBINE pairs of parents;
3. MUTATE the resulting offspring;
4. SELECT individuals or the next generation

END.

Figure. 3

In GAs, the individuals/chromosomes are encodedin the form of strings andcollection of such strings called a *population*. Initially, arandom population is created, which represents differentpoints in the search space. An *objective* and *fitness*function is associated with each string that represents thedegree of *goodness* of the string. Based on the principle of survival of the fittest, a few of the strings are selected more than one time andsome of them are not selected even once for themating pool. Biologically inspired operators

likecrossover and *mutation* are applied on these strings toyield a new generation of strings. The process ofselection, crossover and mutation continues for a fixednumber of generations or till a termination condition issatisfied.

4. Overview of GA based clustering algorithms and their applications

The process of grouping a set of physical or abstract objects into classes of *similar* objectsis called clustering. The cluster analysis represents a group of methods whose aim is to classify theinvestigated objects into clusters. The founders of cluster analysis were Tryon, Wardand James.One of the majorproblems encountered by researchers while using different clustering methods is that each method can generate different solutions for the same data. Due to this problem we need algorithms which can discover themost `natural' groups in a data set.

Krovi R. made a research to investigate thefeasibility of using genetic algorithms for theclustering task of data mining [6].

K.Krishna and M.N.Murty proposed a novel hybrid genetic kmeansalgorithm (GKA) [7] to find a globally optimal partition of a given data into a specified number of clusters. The proposed GA circumvent expensive crossover operator used to generate valid child chromosomes from parent chromosomes. It hybridized the GA by using a classical gradient descent algorithm used in clustering viz., K-means algorithm. In *genetic Kmeansalgorithm* (GKA), K-means operator was definedand used as a search operator instead of crossover. It defined a biased mutation operator specific to clustering called distance-based-mutation. The authors used finite Markov chain theory to prove that the proposed GKA converges to the global optimum. It was also observed that GKA searches faster than some of the other evolutionary algorithms used for clustering.

An improved version of GKA known as *Fast Genetic K-means Algorithm (FGKA)* was proposed in [8]. The proposed GA featured several improvements over GKA. It was evident from experiments in [7][8] that K-means algorithm might converge to a local optimum, both FGKA and GKA always converge to the global optimum. FGKA initializes the population to P_0 and obtains the next population by applying selection, crossover and mutation operators and it keeps on evolving until some termination condition is met. Illegal strings are permitted in FGKA during initialization phase, but were considered as the most undesirable solutions by defining theirtotal within cluster variation (*TWCV*s) as infinity ($+\infty$). By allowing illegal strings the overhead of illegal string in the evolution process was avoided and thus improved the time performance of the algorithm as compared to GKA.

Incremental Genetic K-means Algorithm (IGKA) proposed in [9] was an extension to previously proposed clustering algorithm, the Fast Genetic K-means Algorithm (*FGKA*). The performance of IGKA was found to be better when the mutation probability was small. IGKA was based calculating the Total Within-Cluster Variation (TWCV) and to cluster centroids incrementally whenever the mutation probability was small for the clustering task. Like FGKA, IGKA also always converges to the global optimum.

A GA-based clustering approach was proposed in [10], which determined k i.e. the number of clusters to be formed by itself.The sum of the Euclidean distances of the points from their respective cluster centers was adopted as the clustering metric in [10].

A GA-based unsupervised clustering technique was proposed in [11], which selects cluster centers directly from the data set, thus speeding up the fitness evaluation process by constructing a look-up table in advance and saving the distances between all pairs of data points. Binary representation rather than string representation is used to encode a variable number of cluster centers and more effective operators for selection, crossover, and mutation were introduced. Finally, the Davies-Bouldin index [12],[13] was employed in the above algorithm to measure the validity of clusters. The new enhanced algorithm has shown a more stable clustering performance.

A GA for automatically evolving the number of clusters as well as doing proper clustering of any data set was proposed in [14]. The individual representation is different from the one used in earlier GAs. It comprised both real numbers and the do not care symbols, to encode a variable number of clusters.The Davies–Bouldin index was used as a measure of the validity of the clusters. Effectiveness and utility of the genetic clustering scheme was demonstrated for a satellite image of a part of the city Calcutta. The proposed technique was able to distinguish some characteristic land cover types in the image in [15].

A novel clustering algorithm for mixed data was proposed in [16].Most of the existing clustering algorithms were only efficient for the numeric data rather than the mixed data set but the proposed GA worked efficiently for datasets with mixed values by modifying the common cost function.

A hybrid genetic based clustering algorithm, called *HGA-clustering* was proposed in [17] to explore the proper clustering of data sets. This algorithm, with the

cooperation of tabu list and aspiration criteria, has achieved harmony between population diversity and convergence speed.

A genetic algorithm was proposed in [18] which designed a dissimilarity measure, termed as Genetic Distance Measure (GDM) to improve the performance of the K-modes algorithm which is an extension of k-means.

A hybrid genetic based clustering algorithm calledSPMD (Single Program Multiple Data) was presented in [19] to improve the convergence of GA as well as to accelerate the convergence speed of GA. It is combination of GA with local searchingalgorithm – uphill. The SPMD algorithm exploitsthe parallelism of GA and at the same time overcomes thepremature and poor convergence properties of GA. Thealgorithm was applied on typical multiple local minimafunctions, TSP problem and an engineering computationproblem QCBED on author developed cluster systemTHNPSC-1.

G*enetic Weighted K-means Algorithm* (GWKMA) proposed in [20] combines genetic algorithm (GA) with a weighted K-means algorithm (WKMA). GWKMA encodes each individual by a partitioning table which uniquely determines a clustering, and employs an extra operator, WKMA operator along with three operators (selection, crossover, Mutation).

A hybrid GA-based clustering (HGACLUS) algorithm proposed in [21] combines genetic algorithm (GA) with simulated annealing to find optimal solution. This algorithm maximized the clustering success by achieving internal cluster cohesion and external cluster isolation. The performance of HGACLUS was compared with other existing clustering methods and was found to be more accurate and robust than other methods.

An improved genetic algorithm (IGA) was proposed in [22] in which an efficient method of crossover and mutation were implemented. The proposed algorithm was a combination of GA, the popular Nelder-Mead(NM) Simplex search and K-means to find optimal solution.

A genetic based clustering algorithm called GASDCA (GA with point Symmetry Distance based Clustering Algorithm) was proposed in [23] which was able to detect both convex and non convex clusters. The proposed GASDCA was compared with existing symmetry based clustering technique, SBKM, its modified version, Mod-SBKM and the well known K-means algorithm [23].

Clustering is one of the most important clustering tasks due to its large area of application. However, one of the main difficulties associated with clustering is the validation of the results obtained as same/different clustering algorithms yield different results on the same dataset. Traditional clustering algorithms were combined with a Genetic Algorithm (GA) to build clusters that gave better real distribution of the datasets. The GA employs a fitness function that combines two validation criteria. Such combination allows the GA to improve the evaluation of the candidate solutions. Furthermore, this combined approach avoids the individual weaknesses of each criterion.

GA based clustering techniques have a large area of application. A few of these applications are discussed in this paper. An evolutionary fuzzy clustering method with knowledge-based evaluation was proposed in [24] to identify unknown functions of genes.

The image compression problem using genetic clustering algorithms based on the pixels of the image was proposed in [25]. GA was used to obtain an ordered representation of the image and then the clustering was performed to obtain the compression.

A GA was proposed in [26] to deal with document clustering. The GA algorithm calculated the optimum value of k and solved thebest grouping of the documents into these k clusters. The performance of this algorithm was evaluated on datasets of documents thatwere the output of a query in a search engine.

A GA was proposed and was applied to enhance the performance of clustering algorithms in mobile ad hoc networks in [27].

5. Conclusion

This paper presents a review on genetic algorithms proposed for the clustering task of data mining. It is evident from the review done that GAs are highly capable of performing successful clustering and their capabilities of GAs were applied for evolving theproper number of clusters and providing appropriateclustering. Many GA based clustering algorithms arestudied. The GAs proposed so far have been applied to different types of datasets, small data setsas well as large data sets, simple datasets as well as multivariate datasets. GA based clustering techniquescan be used in many application areas like document clustering,image compression, gene expression analysis and textclustering etc. GA was applied on Clustering algorithmslike K-means and fuzzy c-means which are mostlydistance based clustering algorithms. GA is yet to beapplied to other clustering algorithm.

6. References

[1] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[2] A. A. Freitas, “A survey of evolutionary algorithms for data mining and knowledge discovery,”

in *Advances in evolutionary computing*, Springer, 2003, pp. 819–845.
[3] A. A. Freitas, *Data Mining and Knowledge Discovery with Evolutionary Algorithms*.Springer, 2002.
[4] A. A. Freitas, "A review of evolutionary algorithms for data mining," in *Soft Computing for Knowledge Discovery and Data Mining*, Springer, 2008, pp. 79–11
[5] Z. Michalewicz, *Genetic algorithms+ data structures= evolution programs*. springer, 1996.
[6] Krovi, R, "Genetic algorithms for clustering: a preliminary investigation", System Sciences, Proceedings of the Twenty- Fifth Hawaii International Conference, Volume: iv On page(s): 540-544, Date: 7-10 Jan 1992,
[7]K. Krishna and M. N. Murty, "Genetic K-Means Algorithm", *IEEE Transaction On Systems, Man, And Cybernetics—Part B:CYBERNETICS*, Vol. 29, No. 3, June 1999
[8] Yi Lu, Shiyong Lu, FarshadFotouhi ,"FGKA: A Fast Genetic K-means Clustering Algorithm", *SAC'04* Nicosia, Cyprus. , March 2004 ACM 1-58113-812-1/03/04
[9] Yi Lu1, Shiyong Lu1, Farshad Fotouhi1, Youping Deng, d. Susan, J. Brown," an Incremental genetic K-means algorithm and its application in gene expression data analysis", *BMC Bioinformatics* 2004
[10] U. Maulik, S. Bandyopadhyay," Genetic algorithm-based clustering technique", Pattern Recognition 33, 2000
[11] H.J. Lin, F.W. Yang and Y.T. Kao," An Efficient GAbased Clustering Technique", Tamkang Journal of Science and Engineering, Vol. 8, No 2, pp. 113_122, 2005
[12]Davis, D. L. and Bouldin, D. W., "A Cluster Separation Measure," *IEEE Trans. Pattern Analysis and MachineIntelligence*, Vol. 1, pp. 224-227,1979.
[13] Bezdek, J. C., "Some New Indexes of Cluster Validity," *IEEE Trans. Systems, Man, and Cybernetics - Part B:*
[14]Bandyopadhyay, S. and Maulik, U., "Nonparametric Genetic Clustering: Comparison of Validity Indices," in *IEEETrans. Systems, Man, and Cybernetics – Part C: Applicationand reviews*, Vol. 31, pp. 120-125, 2001
[15]Bandyopadhyay, S. and Maulik, U., "Genetic Clustering for Automatic Evolution of Clusters and Application to Image Classification,"*PatternRecognition*,Vol.35,pp.1197-1208, 2002
[16] LI Jie, G. Xinbo, "A GA-Based Clustering Algorithm forLarge Data Sets With Mixed Numeric and Categorical Values*",IEEE, Proceedings of the Fifth International Conference onComputational Intelligence and Multimedia Applications(ICCIMA'03)* 0-7695-1957-1/03, 2003
[17] Y. Liu, Kefe and X. Liz," A Hybrid Genetic Based Clustering Algorithm*", Proceedings of the Third InternationalConference on Machine Leaming and Cybernetics*, Shanghai,26-29 August 2004
[18] S. Chiang, S. C. Chu, Y. C. Hsin and M. H. Wang, "Genetic Distance Measure for K-Modes Algorithm",*International Journal of Innovative Computing, Informationand Control ICIC,ISSN 1349-4198, Volume 2, Number 1*, pp.33-40 February 2006
[19] Zhihui D., Meng D., Sanli Li, Shuyou Li, Mengyue Wu and Jing Zhu, "Massively Parallel SPMD Algorithm for Cluster Computing: Combining Genetic Algorithm with Uphill.
[20] Fang-Xiang Wu, Anthony J. Kusalik and W. J. Zhang, "Genetic Weighted K-means for Large-Scale Clustering Problems", University of Saskatchewan, CANADA
[21] H. Pan, J. Zhu, DanfuGeno., " Genetic Algorithms Applied to Multi-Class Clustering for Gene Ex-pression Data", Geno., Prot. &Bioinfo. Vol. 1 No. 4 November 2003
[22] V. Katari, S. C. Satapathy, JVR Murthy,P. Reddy ,"AHybridized Improved Genetic Algorithm with Variable LengthChromosome for Image Clustering", *International Journal ofComputer Science and Network Security*, VOL.7 No.11,November 2007
[23]S. Saha, S. Bandyopadhyay, U. Maulik," A NewSymmetry-Based Genetic Clustering Algorithm", *MachineIntelligence Unit, Indian Statistical Institute, India*
[24]Han-Saem Park, Si-Ho Yoo, and Sung-Bae Cho "Evolutionary Fuzzy Clustering Algorithm with Knowledge-Based Evaluation and Applications for Gene Expression",*Journal of Computational and Theoretical Nanoscience*Vol.2,1–10, 2005
[25] Merlo, Caram, Fernández, Britos, Rossi, &García-MartínezR.,"Genetic-Algorithm Based ImageCompression",SBAI–SimpósioBrasileiro de AutomaçãoInteligente, São Paulo, SP, 08-10 de Setembro de 1999
[26] A. Casillas, M. T. Gonzalez de Lena, and R. Martınez,"Document Clustering into an unknown number of clusters usinga Genetic Algorithm"
[27]DamlaTurgutSajal K. Das, RamezElmasriandBegumhan,"OptimizingClusterinAlgorithm inMobile Ad hocNetworks Using GA approach