International Journal of Pure and Applied Mathematics
Volume 118 No. 24 2018
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

A Review on Clustering using Evolutionary Computation

Pravati Sahoo[1],Dr.Subrat Kumar Nayak[2]
Dept. Of CSE
SOA Deemed to be University
Bhubaneswar, India
pravati.sahoo728@gmail.com,
subratnayak@soa.ac.in ,

May 23, 2018

Abstract

Important role of data analysis for extracting various phenomena starts with clustering techniques. Starting fromK-means (1957) algorithm, many partitioned clustering algorithms have been extensively analysed and reported. The availability of many approaches may confuse a researcher to adopt one approach out of many. In general, most of the researchers have focused on the convergence property of the proposed approach. The convergence issue of many traditional algorithms has been nicely handled with the help of nature inspired metaheuristic optimisationtechnique. Clustering, being a data mining approachcan be considered as an optimisation problem.Hence, nature inspired optimisation algorithmsplay a distinguish role for extracting knowledge from a dataset. This paper reviews different metaheuristic approaches and their applications in clustering problem to solve major issues.

Key Words:Data mining, clustering, evolutionary computation .

1 Introduction

Grouping the samples in relevant categories,can be considered as a key issue for many applications. Basically, data clustering [1] ensures a preliminary structure for large database. This helps in faster knowledge extraction and abstract interpretation of samples. The categories must be wellrepresented, inspite of the dependencies among the variables are not completely understood. The process can be carried out either in supervised or unsupervised manner. The classification exploits the known target classes associated with the training samples, whereas theclustering defines the groups of patterns based on the analysis performed on the dataset. One of the most common applications of classification/clustering is pattern recognition [2]. In this case, each cluster is associated to a category of patterns. So once a sample is included into a cluster, that sample is interpreted as an instance of the corresponding category and linked with a specific concept.

Because of the nonlinear landscape in input space and the huge number of features characterizing each sample, the categorisation of a dataset into number of clusters is a challenging process. Obviously, the problem becomes even more difficult when working with a large and unbalanced dataset (i.e. containing very different number of samples in each cluster and/or samples non-uniformly distributed within each group).

This paper aims to provide an insightful overview of clustering algorithms. In Section II basic anatomy is reviewed, which outlines the recent trends in the field and focuses the main clustering algorithms. Different evolutionary algorithms for clustering is briefly discussed in Section III. Finally, several summarizing remarks are presented in section IV.

2 ANATOMY OF CLUSTERING

This section briefly describes how clusteringcan be formulated as an optimization problem and overview of the background studies.

A.Problem Formulation

Considering an unlevelled dataset X= [X1, X2, X3.Xn]n*d. Here,n represents number of patterns andd is the number of features.Clustering technique aims to cluster the dataset into K groups

(K¡n), such thatCk for all k=1,2K.Here, Ckis the kth cluster.Clustering strongly depends on similarity of data present in a dataset.If we frame as an objective function for clustering task for optimisation thenoptimizeCk [(X,Ck)] for all k=1,2,K.

B. Background

With the evolution of number of metaheuristic approaches,the process of optimization even become simpler for complex objective function. Here, an optimal solution can be achieved easily by considering the entire search space along with its adaptive search steps.

Some of the widely used clustering techniques are K-means algorithm [3], maximum likelihood estimate method [4], graph theoretic approaches [5] and branch and boundtechnique [6]. K-means algorithm, is one of the most extensively used algorithms, which can efficiently solve a clustering problem. However, K-means suffers from a lacuna of being trapped in local minima because of its initial step. A tree search technique is used in branch and bound procedure. The speciality of this technique is to remove the subtrees that may not contain the optimal outcome. But the challenge begins with a large dataset, where deciding the criteria for eliminating sub-tree is quite difficult [7]. Later, assuming a specific distribution of a huge dataset, probabilities of the classes are computed in maximum likelihood prediction technique. Using graph theory, cluster can be formed by constructing direct tree from data by approximating the compactness of gradient at every point. The valley of the density function can be realised as a cluster. Specifically, in the low-density area, the quality of the outcome relies completely on the quality of the estimation or prediction procedure for the density gradient.

Many of the earlier articles have been proposed for clustering technique, which have not adopted any known distribution of the given dataset. On the other hand, to be a good algorithm, an algorithm should not suffer from the limitation to provide imperfect optimal clustering based on the initial choice of clusters. To overcome the above-mentioned limitations of the traditional algorithms, evolutionary algorithms plays an important role in finding the optimal solution.

3 EVOLUTIONARY ALGORITHMS

This section represents clustering byusing different evolutionary techniques.

A. Genetic Algorithm

Genetic algorithms (GAs) [8] are optimization techniques based on randomized search influenced by the ethics of evolution and natural genetics. It has been proved that GA can handlecomplex, large range and variance search space in an effective manner and results good output for objective or fitness function for different optimization problem. Limits of the entire exploration domain are represented in the form of collection of genes in GA. A group of such genes is called a population. At first, anarbitrary population is formed between desired search range. An objective and fitness function are applied on each population that shows the step of goodness of each population. Based on the opinion of Darwins theory of the fittest, some better populations participate in the mating pool for next generation. Biologically inspired operators like crossover and mutation are used on these chromosomes to develop new generation of chromosomes.

GA based approaches

Initially, Bezdek et al. [9] proposed the use of basic genetic algorithm for partitional clustering. The author describes an approach to optimize the clusters, which has been created during unsupervised clustering by using genetic algorithm. Chromosome initialization is being done by the standard binary encoding scheme with xed number of cluster centres[1011].Cluster-oriented mutation and uniform crossover these two are used for reproduction operation.Subsequently, Murthy and Chowdhury [12] proposed integer-based encoding of chromosomes.In this paper, GA have been used to optimize a specified objective function which is related to a clustering problem.The use of real coded genetic algorithm for partitional clustering [13] have been proposed by Maulik and Bandyopadhyay, where the computational complexity is reduced to O(k) compared to O(nk), which is associated with integer or binary encoding. GA is used for finding the appropriate cluster centresfor which the metric is optimized. By Krishna and Murty [14], a genetic K-mean algorithm have been proposed. In this, the crossover phenomenon is substituted with K-mean by the basic search operation.

B. Particle Swarm Optimisation

In 1995, the Particle swarm optimisation (PSO) algorithm proposed by Kennedy and Eberhart as a random population-based optimization process [15]. It is stimulated from the behaviour of the individuals (particles) in a bird flock or fish school, where particles show a natural behaviour of tracing the target distance when they are in search of some target (e.g., food) [16]. The aim is to reach the global best of some multidimensional and complex nonlinear function that helpsin updating the position of information recovered from the local best and global best. So, PSO follows almost same strategy as of other evolutionary algorithms (EAs), like genetic algorithm (GA). For updating the position out of three parameters, its inertia or momentum, which serves as a memory of the previous particle direction that plays a crucial role in preventing the particle from suddenly changing direction.

PSO based approaches

Van der Merwe and Engelhrecht [17] applied the above-mentioned approach in cluster analysis of arbitrary datasets.Particle swarm clustering (PSC) algorithm have been proposed by Cohen and Castro [18]. With the help of cognitive term, social term and self-organizing term the previous particles are accessed to update the velocity of particles and it helps the particle to avoid the local stagnation.Omran et al. [19] proposed the cluster analysis for image clustering using PSO. In this paper, PSO is used for finding the accurate centroids of a user specified number of clusters and examined the application of the PSO to cluster data vectors. Cura et al. [20] recently developed a new PSO based partitional clustering algorithm to handle the unidentified number of clusters. In this paper, the clustering problem can be successfully solved for both known and unknown numbers of clusters. Zhang et al. [21]proposed a model to combine PSO with possibility C-means(PCM) for image segmentation which provides better performance as compare to the fuzzy C-means(FCM) algorithm.

C. Ant Colony Optimization

The Ant colony optimization (ACO)[22] algorithm is inspired by ants behaviour which determines the optimal path from the nest to the food source. Ant-based clustering and sorting algorithm was initially presented for the task in robotics by Deneubourget al. But Lumer and Faieta extended this application of the proposed al-

gorithm for numerical data analysis. Hence, the proposed model is very efficiently used for data-mining algorithms. The modified ant-based methods have proved their goodness and effectiveness in many test cases. However, earlier ant-based clustering approach was generally not matured and had many issues for enhancement. Having these predictions, sometimes, the typical ant-based clustering can perform well. The foremost cause of refinement isthat the clusters obtained from these clustering algorithms failed to achieve 100% accuracy. The errors known as misclustering because there is chanceof selecting wrongclustering parameter. Refinement algorithm can be used to avoid such errors. In some of the methods, only one ant can be used for refinement of clusters. By allowing this antmove for a random walk on the selected clusters.

ACO based approaches:

Initially, Ants based K-means algorithm have been proposed by Kuo et al. [23] i.e. subsequently upgraded by grouping of self-organizing maps(SOM), ACOand k-means in [24]. A variant of ACO, known as APC (aggregation pheromone density-based clustering) algorithm is proposed by Ghosh et al. [25]. The APC is used for updating the pheromone matrix which is helpful to avoid the convergence of a solution.Wan et al. [26] proposed a modified algorithm known as chaotic ant swarm (CAS). In this paper, CAS gives optimal partitions irrespective of cluster size and density. Yang et al. [27] proposed the use of multi-ant colonies algorithm for clustering.In this concept, several independent ant colonies (each having a queen ant) are presented. Clustering results are produced from each colony in parallel and sends it to the queen ant agent. Recently, a new approachhas been proposed by Zhang and Cao [28] by hybridizing kernel principal component analysis (KPCA) with ACO. In this paper, the efcient features on the dataset is calculated by KPCAand then in the feature space, the ant-based clustering is performed.

D. Artificial Bee Colony

Karaboga proposed Artificial bee colony (ABC) algorithm for optimizing numerical problems in [25]. The algorithm inspired the intelligent foraging behavior of honey bee swarms. It is a very simple, robust and population based stochastic optimization algorithm. The colony of artificial bees comprises of three collections of bees: employed bees, onlookers and scouts. A bee waiting on the dance

area to select a food source is called onlooker and one going to the food source visited by it before is named as employed bee. Then the scout bee is carried out arbitrary search for discovering new sources. The position of a food source denotes a possible solution to the optimization problem and the nectar amount of a food source represents to the quality of the solution.

ABC based Approaches:

A modified version of ABC named as Chaotic ABC has been introduced by Zhang et al. [30], to solve the partitional clustering problems. In this algorithm, initialization of parameters in Rossler chaotic number generator was combined to ensure better quality of solutions and more robustness. Yan et al. [31] proposed a hybrid clustering algorithm HABC by integrating crossover operation of GA in ABC. The result of thisalgorithm provides betterperformance as compared to PSO, CPSO, GA, ABC and K-means algorithms.ABC was modified by Zou et al. [32] which employing the cooperative strategy to find better solution with involvement of each individual.In k-Means clustering, ABC is used by Ju et al. [33] to reduce the local optimal problems and produceefficient clusters. By using traditional k-modes technique, Ji et al. [34] developed ABC-K-Modes clustering algorithm for categorical data. For solution search in employed and onlooker phase, One-step k-modes procedure isbeing used.

4 CONCLUSION AND FUTURE SCOPE

An up-to-date analysis of nature inspired algorithms for partitional clustering is done. It is observed that the traditional gradient based partitional algorithms are computationally simpler but often provide inaccurate results as the solution is trapped in the local minima. Use of entire search space and faster search algorithms ensures optimal cluster formation. These promising solutions of metaheuristic approaches are possible because of searching entire space along with better updating formulations even though for unlabelled data set. Every clustering algorithm has its own impact on a specific application area.

Some of the optimisation techniques have shown their ability of quick convergence where some shown for best convergence result.

Many researches are also going on for hybridising the behaviour of different optimisation algorithms to improve its robustness for different applications.

5 Acknowledgement

Author would like to acknowledge Dr.Subrat Kumar Nayak (SOA Deemed to be University) for his guidance and support.

References

[1] Chitra, K., &Dr., Maheswari, D. (2017). A Comparative Study of Various Clustering Algorithms in Data Mining. pp. 109-115, IJCSMC, Vol. 6, Issue. 8.

[2] Duin, R. P., &Pekalska, E. (2005). Open issues in pattern recognition. In Computer Recognition Systems (pp. 27-42). Springer, Berlin, Heidelberg.

[3] Gonzlez, R. C., &Tou, J. T. (1974). Pattern recognition principles. Applied Mathematics and Computation. Reading (MA): Addison-Wesley.

[4] Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 5(3), 329-350.

[5] Koontz, W. L. G., Narendra, P. M., &Fukunaga, K. (1976). A graph-theoretic approach to nonparametric cluster analysis. IEEE Transactions on Computers, (9), 936-944.

[6] Koontz, W. L. G., Narendra, P. M., &Fukunaga, K. (1975). A branch and bound clustering algorithm. IEEE Transactions on Computers, 100(9), 908-915.

[7] Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, Academic Press, New York.

[8] Ribeiro Filho, J. L., Treleaven, P. C., &Alippi, C. (1994). Genetic-algorithm programming environments. Computer, 27(6), 28-43.

[9] Bezdek, J. C., Boggavarapu, S., Hall, L. O., &Bensaid, A. (1994, June). Genetic algorithm guided clustering. In Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on (pp. 34-39). IEEE.

[10] Sarkar, M., Yegnanarayana, B., &Khemani, D. (1997). A clustering algorithm using an evolutionary programming-based approach. Pattern Recognition Letters, 18(10), 975-986.

[11] Kuncheva, L. I., &Bezdek, J. C. (1997). Selection of cluster prototypes from data by a genetic algorithm. In Proc. 5th European congress on Intelligent techniques and soft computing (pp. 1683-1688).

[12] Murthy, C. A., & Chowdhury, N. (1996). In search of optimal clusters using genetic algorithms. Pattern Recognition Letters, 17(8), 825-832.

[13] Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. Pattern recognition, 33(9), 1455-1465.

[14] Krishna, K., &Murty, M. N. (1999). Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(3), 433-439.

[15] Ghorpade-Aher, J., &Bagdiya, R. (2014). A Review on Clustering Web data using PSO. International Journal of Computer Applications, 108(6).

[16] Esmin, A. A., Coelho, R. A., &Matwin, S. (2015). A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artificial Intelligence Review, 44(1), 23-45.

[17] Van der Merwe, D. W., & Engelbrecht, A. P. (2003, December). Data clustering using particle swarm optimization. In Evolutionary Computation, 2003. CEC'03. The 2003 Congress on (Vol. 1, pp. 215-220). IEEE.

[18] Cohen, S. C., & de Castro, L. N. (2006, July). Data clustering with particle swarms. In Evolutionary Computation, 2006. CEC 2006. IEEE Congress on (pp. 1792-1798). IEEE.

[19] Omran, M. G., Engelbrecht, A. P., & Salman, A. (2004). Image classification using particle swarm optimization. In Recent Advances in Simulated Evolution and Learning (pp. 347-365).

[20] Cura, T. (2012). A particle swarm optimization approach to clustering. Expert Systems with Applications, 39(1), 1582-1588.

[21] Zhang, Y., Huang, D., Ji, M., &Xie, F. (2011). Image segmentation using PSO and PCM with Mahalanobis distance. Expert Systems with Applications, 38(7), 9036-9040.

[22] Yang, X. S. (2010). Nature-inspired metaheuristic algorithms. Luniver press.

[23] Kuo, R. J., Wang, H. S., Hu, T. L., & Chou, S. H. (2005). Application of ant K-means on clustering analysis. Computers & Mathematics with Applications, 50(10-12), 1709-1724.

[24] Chi, S. C., & Yang, C. C. (2006, October). Integration of ant colony SOM and k-means for clustering analysis. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 1-8). Springer, Berlin, Heidelberg.

[25] Ghosh, S., Kothari, M., Halder, A., & Ghosh, A. (2009). Use of aggregation pheromone density for image segmentation. Pattern Recognition Letters, 30(10), 939-949.

[26] Wan, M., Wang, C., Li, L., & Yang, Y. (2012). Chaotic ant swarm approach for data clustering. Applied Soft Computing, 12(8), 2387-2393.

[27] Yang, Y., & Kamel, M. S. (2006). An aggregated clustering approach using multi-ant colonies algorithms. Pattern Recognition, 39(7), 1278-1289.

[28] Zhang, L., & Cao, Q. (2011). A novel ant-based clustering algorithm using the kernel method. Information Sciences, 181(20), 4658-4672.

[29] Karaboga, D., &Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied soft computing, 11(1), 652-657.

[30] Zhang, Y., Wu, L., Wang, S., &Huo, Y. (2011). Chaotic artificial bee colony used for cluster analysis. In Intelligent Computing and Information Science (pp. 205-211). Springer, Berlin, Heidelberg.

[31] Yan, X., Zhu, Y., Zou, W., & Wang, L. (2012). A new approach for data clustering using hybrid artificial bee colony algorithm. Neurocomputing, 97, 241-250.

[32] Zou, W., Zhu, Y., Chen, H., & Sui, X. (2010). A clustering approach using cooperative artificial bee colony algorithm. Discrete Dynamics in Nature and Society, 2010.

[33] Ju, C., & Xu, C. (2013). A new collaborative recommendation approach based on users clustering using artificial bee colony algorithm. The Scientific World Journal, 2013.

[34] Ji, J., Pang, W., Zheng, Y., Wang, Z., & Ma, Z. (2015). A novel artificial bee colony-based clustering algorithm for categorical data. PloS one, 10(5), e0127125.