A Novel Hybrid Candidate Group Search Genetic Clustering for Large Scale Data

Suvarna P. Patil
M. E. Computer Department,
Pune University, PCCOE College,
Pune 44, India

ABSTRACT

Clustering is an unsupervised approach to extract hidden patterns from the datasets. There are certain challenges in clustering, though it is very much difficult to produce good clustering, researchers have provided the solutions through various hybrid approaches. The proposed work is based on enhancing the clustering results by using two algorithms: First Candidate Group Search (CGS) is used to produce clusters and Genetic algorithm (GA). A CGS can be applied to large dataset with less computational time, but the drawback is it can't results in global optima. Hence GA is used for further optimization. Both algorithms will produce optimized clusters.

Keywords

Clustering, Candidate Group Search, Genetic algorithm, global optima.

1. INTRODUCTION

Clustering is one of the most popular techniques in data mining. Clustering can be considered the most important unsupervised learning method; it deals with finding a structure in a collection of unlabeled data. A definition [13] of clustering could be, "the process of organizing objects into groups whose members are similar in some way". Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar in some sense or another to each other than to those groups i.e. clusters. We can show this with a simple graphical example as shown in Figure 1.

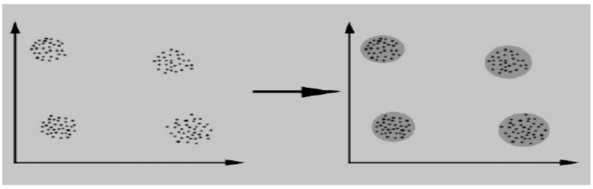

Figure 1 Clustering

The similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance. This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Remaining paper is organized as follows. Section II presents an overview of different clustering techniques and K-Harmonic Mean clustering methods. Section III introduces proposed method. Comparative analysis is done in section IV and conclusion about the paper is made in section V.

2. RELATED RESEARCH WORK

Among all clustering approaches, K-means (KM) might be the most classic and widely used approach due to its merits, simple and fast. First of all, the KM algorithm [3] tries to divide the original data set into K groups to minimize predefined objective function, Min Square Error (MSE). The idea of MSE is to minimize sum of the squared distance between each point to its corresponding center. Keep updating centers until this objecting function converge. But, KM has drawback of convergence to the local optimum.

Because of this Zhang proposed a center-based clustering scheme K-harmonic means (KHM) in 1999 to solve the high sensitivity problem.

The KHM algorithm [3] is also similar to KM only difference is that KHM used different objective function called as harmonic mean. Harmonic mean is used to calculate the group center. Harmonic mean tends to minimize the errors within the group and maximize the error between groups. All centers are replacing randomly or partially using new centers as the initial points, and solved it recursively till to converge.

It is iterative algorithm that refines the clusters defines by K centers. KHM takes the harmonic averages of the squared distance from a data point to all centers as its performance function. The harmonic average of K numbers is defined as-the reciprocal of the arithmetic average of the reciprocals of the numbers in the set.

KHM uses HM to measure the distances between each point and its corresponding center.

The harmonic mean (HM) of N numbers, is defined as,

$$HA = ((a_i \mid i=1\ldots K)) = \frac{K}{\sum_{i=1}^{K}\frac{1}{ai}} \qquad \ldots\ldots..\ (1)$$

But KHM algorithm requires more computational time for large scale data.

A Candidate group search (CGS) [3, 5], also uses KHMs algorithm to obtain the center of each groups. Based on KHM's solutions, CGS perturbs the solution, replacing centroid, to escape from local optimum. By taking each center as a core and screening all the data points according to the ratio of the distance between points to the core and the maximum distance, we can find each candidate group for current centroid. Use this candidate group as the neighbour set and choose possible candidate to replace current center. Screening through all the data set if it fit in to the candidate group, the center has to be interchanged and using new solution is achieved by using KHM. Candidate Group Search offers a scheme combining of some haphazardness and deterministic selection rules coming from the data set, CGS outperforms than KHM and it requires less computation time.

A genetic algorithm (GA) [8] may be described as a mechanism that imitates the genetic evolution of species. GA starts with a population of chromosomes. Each chromosome is evaluated based on the problem's objective function and is given a fitness value. The chromosome with better fitness value has a higher probability of being selected to participate in the following crossover operation. The purpose of the crossover operation is to exploit promising areas of the search space. Through this operation, GA combines parts of chromosome pairs (parents) to produce new chromosomes (offspring). The offspring seek to inherit the advantages of both parents. The purpose of another G A operation, known as mutation, is to explore new regions of the search space that may not be represented by any of the current population members. This operation involves altering a small percentage of randomly selected chromosomes in the population. Through a combination of these two (and sometimes other) operations, a GA attempts to produce offspring chromosomes whose fitness values are better than those of their parent chromosomes.

3. PROPOSED WORK

A novel hybrid Candidate Group Search Genetic Clustering algorithm is proposed to achieve global optima in clustering of large scale data. A CGS can be applied to large dataset with less computational time, but the drawback is it can't results in global optima. Hence GA is used for further optimization. Both algorithms will produce optimized clusters.

The main idea of CGS algorithm is using the concept of candidate group set. First, CGS use some selection rules to identify the candidate group set for each center. Screening through all the data set, if it belongs to the candidate group, the center has to be replaced and using KHM to obtain a new solution. CGS provides a scheme combining of some randomness and deterministic selection rules coming from the data set. We found that CGS get better performance in KHM's objective function and require less computational time.

Figure 2 shows the complete flow of the proposed model.

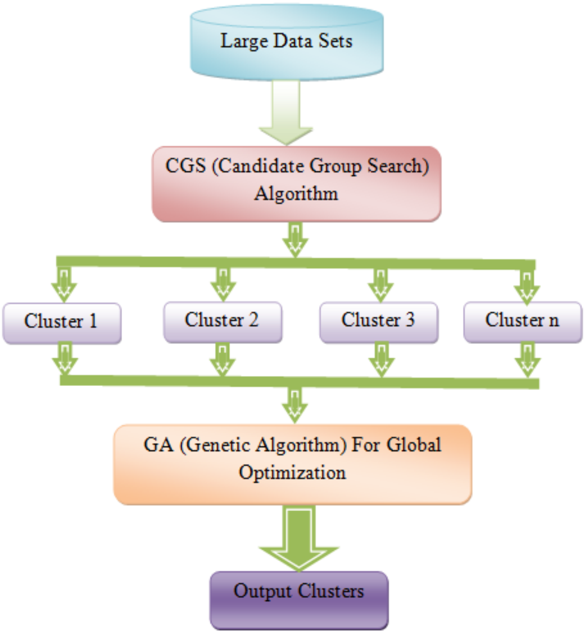
Large Data Sets
CGS (Candidate Group Search) Algorithm
Cluster 1
Cluster 2
Cluster 3
Cluster n
GA (Genetic Algorithm) For Global Optimization
Output Clusters

Figure 2 Architecture of Proposed Work

4. COMPARATIVE STUDY

Following is the Table 1 showing comparative analysis of the different clustering algorithm on the basis of different criteria.

Table 1 Comparative Study

Characteristic	Achieve Local/Global Optima	Computational Time	Applicable to Large Dataset	Implementation	Clustering Shape
Candidate Group Search Algorithm	Local Optima	Required Less Computational Time	Applicable to large dataset	Easy	Arbitrary
Genetic Algorithm	Global Optima	Required More Computational time if dataset is large	Applicable to large dataset	Easy	Non Spherical
Gravitational Search Algorithm	Global Optima	Required More Computational time But Superior than KHM	Applicable to large dataset	Easy	Hyperspherical
K-Harmonic Means Algorithm	Local Optima	Required More Computational Time	Not Applicable to large dataset	Easy	Arbitrary
K Means Algorithm	Local Optima	Required More Computational Time	Not Applicable to large dataset	Easy	Spherical

5. CONCLUSION

In this work following is the conclusion:

Here, we propose a new searching scheme, a novel hybrid Candidate Group Search Genetic Clustering algorithm. The candidate view is based on the distance between the centroid to all entities. By taking each center as a core and defining possible candidate group, CGS can reduce the computational time since the searching size is small. By capability of Genetic algorithms we find accurate, optimally disjoint partitions and proper number of clusters for a dataset. Using the hybrid approach of CGS and GA clustering algorithm we lead to global optimal solution.

6. REFERENCES

[1] H. Jiang, S. Yi, J. Li, F. Yang, X. Hu, “Ant clustering algorithm with k-harmonic means clustering”, 2010.

[2] Hua Jiang, Shenghe Yi, Jing Li, Fengqin Yang, Xin Hu, “Ant clustering algorithm with K-harmonic means clustering”, 2010.

[3] Cheng Huang Hung, Hua-Min Chiou, Wei-Ning Yang, “Candidate groups search for K-harmonic means data clustering”, Applied Mathematical Modelling, Elsevier, 2013.

[4] Habiba Drias, Ilyes Khennak, Anis B, “A Hybrid Genetic Algorithm for large scale Information Retrieval”, 2009.

[5] http://research.ijcaonline.org/volume88/number17/pxc3894002.pdf, “Introducing Hybrid model for Data Clustering using K-Harmonic Means Gravitational Search Algorithms”, 2014.

[6] Zulal Gungor, Alper Unler, “K-Harmonic means data clustering with tabu-search method”, Applied Mathematical Modelling, Elsevier, 2007.

[7] YusenLi a, JunYu b, DapengTao, “Genetic algorithm for spanning tree construction in P2P distributed interactive applications”, Neurocomputing, Elsevier, 2014.

[8] Yuan Chen, Zhi-Ping Fan, Jian Ma, Shuo Zeng, “A hybrid grouping genetic algorithm for reviewer group construction problem”, The Expert Systems with Applications, Elsevier, 2011.

[9] Abdolreza Hatamlou, SalwaniAbdullah, HosseinNezamabadi-pour, “A combined approach for clustering based on K-means and gravitational search algorithms”, Swarm and Evolutionary Computation, Elsevier, 2012.

[10] Kweku-Muata, Osei-Bryson, “Towards supporting expert evalution of clustering results using a data mining process model”, 2009.

[11] Yi Hong, Sam Kwong, “To combine steady-state genetic algorithm and ensemble learning for data clustering”, 2008.

[12] Jiawei Han, By Han Kamber, “Data Mining Concept and Techniques”, 2nd Edition.

[13] Ashok Kumar Thavani Andu, Dr Antony Selvdoss Thanamani, “Multidimensional Clustering Methods of Data Mining for Industrial Applications”, 2013