Introducing Hybrid Model for Data Clustering using K-Harmonic Means and Gravitational Search Algorithms

Anuradha D. Thakare
Asst. Professor
Computer Engineering Dept.
PCCOE, Pune.

Rohini S Hanchate
ME Scholar
Computer Engineering Dept.
PCCOE, Pune.

ABSTRACT

Clustering is a process of extracting reliable, unique, effective and comprehensible patterns from database. Various clustering methods are proposed to accomplish exactness and accuracy of clusters. K-Means is well known clustering algorithm but it easily converge to local optima. To overcome this drawback, an improved algorithm called K-Harmonic Mean (KHM) was proposed, which is independent of cluster center initialization. This article presents study of hybridization KHM with other clustering algorithms. In order to improve the clustering accuracy the authors proposed new hybrid KHM model.

Keywords

K-Harmonic Mean, Clustering algorithm, Genetic Algorithm.

1. INTRODUCTION

Motivation behind clustering is to find an inherent structure in the data, the data objects within each cluster group should exhibit large degree of similarity while the similarity among different clusters should be minimized. Clustering is one of essential processing step which extracts data from huge data warehouse and creates data patterns so that the similar patterns can be grouped.

K-means clustering is center based approach which arbitrarily selects the k number of clusters of the objects in data set each of which primarily denotes a cluster center. Based on Euclidean distance between the object and mean value it assigns objects to the cluster where the objects are highly similar. The k-means algorithm is iterative process and for all the clusters it computes the new mean using the objects dispensed to the cluster in the earlier iteration. By using update mean value as the new cluster centers reassigns all the objects [1].

Gravitational search algorithm (GSA) is newly developed method used for optimization problem The GSA is built on the law of Newtonian Gravity ‘‘Every object attracts other objects with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them’’. In GSA algorithm, all entities can be observed as objects with masses. The objects attract each other by the gravity force, and gravity force makes object to move towards the object with heavier mass, and object with heavier mass become heavier. The objects transform information by the gravitational force.

Genetic Algorithms (GA) is robust, and its quick searching ability and optimization techniques, they can be used to search the enormous document search space. GA generates the new population in each iteration and calculates the fitness function. After evaluation of fitness function is done, the genetic operator like selection is used which is based on principal of survival of the fittest. Only those chromosomes are selected which satisfies the fitness function for reproduction. The crossover operation produces two new offspring’s. In mutation operation, the main objective is to restore lost and explore variety of data. These operations are applied to compute the new population [2].

Remaining paper is further structured in this way. II section will review different clustering techniques and K-Harmonic Mean clustering methods, comparative analysis is done in section III and conclusion about the paper is made in section IV.

Our goals are to study various clustering techniques based on K-Harmonic Mean and how this clustering method is used to achieve accurate clustering which in turn will be used for fast and efficient information retrieval.

2. RELATED RESEARCH WORK

The numerous clustering approaches are developed to form accurate, effective clusters with less computation time. These existing clustering algorithms can be mainly categorized into the following categories: partitional clustering, Hierarchical clustering, Grid-based clustering method and Density-based clustering. Entire clustering process is as shown in figure 1.

Partitional clustering techniques typically start with the patterns partitioning into a number of clusters and divide the patterns by increasing the number of partitions. Hierarchical clustering operates by partitioning the patterns into successively slighter structures [3].

KHM is a center-based clustering algorithm [4] which uses the Harmonic Averages of the distances from each data point to the centers as it constituent to its performance function.

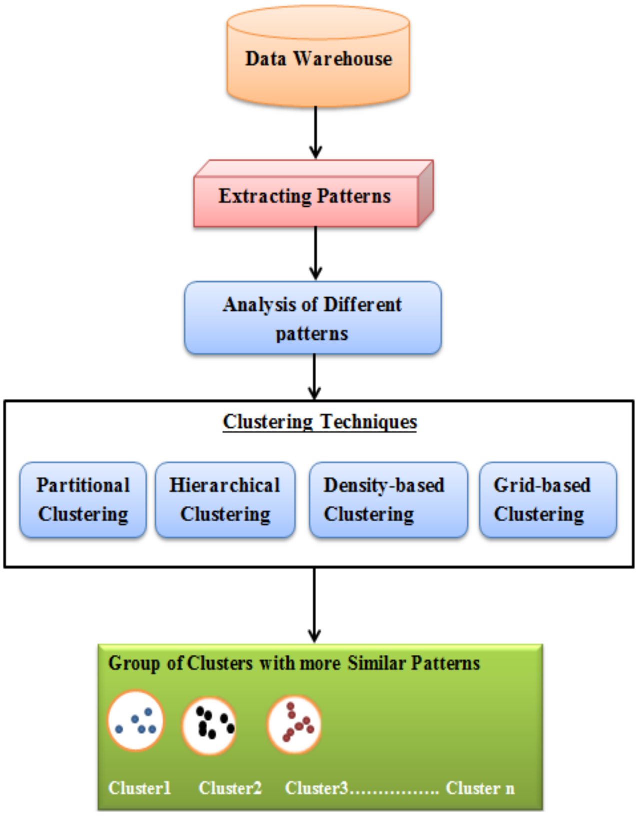

Figure 1: Clustering Techniques

The K-Harmonic Means is essentially oblivious to the initialization of the centers and because of this feature K-Harmonic Mean algorithm overcome the drawbacks of K-Mean. KHM uses different objective functions [4] which are, as follows.

$$KHM = \sum_{k=0}^{n} \left(\frac{\text{K cluster centers}}{\left(\sum \textit{average} \text{ of distance from all points to cluster center}\right)^{p}} \right)$$

In KHM, the objective function computes the membership function and weight associated to each data points. Then it recomputed the centers location from all data points and assigns the data point xi to cluster Ci which is having highest membership value.

Merits of K-Harmonic Mean:

- In some cases K-Harmonic Means significantly improves the quality of clustering results.
- K-Harmonic Means clustering algorithm with a new objective function.
- KHM outperforms than K-Means
- K-Harmonic Means congregates faster than K-Means.
- KHM can be integrated with other local search algorithms.

Drawbacks of KHM:

- K-Harmonic Mean method typically runs into local optima.

GSKHM (Gravitational Search Approach using K-Harmonic Means) method is effective way of clustering the documents and this can be achieved by using combination of Gravitational Search Method and K-Harmonic Means algorithm. The GSA is based on the law of Newton Gravity where each element in the universe attracts every other element with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. In GSA algorithm, all the objects with masses, attracts each other by the gravity force. The force makes all of them move towards the ones with heavier masses. The objects transform information by the gravitational force, and the objects which is having higher mass becomes heavier.

Procedure for GSA:

- Calculate the force acted on object x from object y by using the Euclidian distance between object x and object y.
- Using calculated force on object x by object y calculate the total force acted on the object x.
- Acceleration of object x is calculated, where object having lesser mass they can move with higher speed which is standardly defined by newtons law and is defined as force acted on object divided by mass of object at an instance of time.
- Using Euclidian distance and acceleration, calculate the objects velocity at which speed object is moving and position object with respect to current position of object x.

KHM method inclines to converge faster than the GSA, but usually it gets wedged in local optima. GSA algorithm always converges to better global optima for its admirable global penetrating ability. In this paper, the integration of both GSA and KHM described in section II to form a hybrid clustering algorithm (GSAKHM): by initializing initial parameters and for eight generation the Gravitational Search Method is applied and for four generation K-Harmonic Means algorithm applied .

Advantages of GSAKHM:

- GSAKHM algorithm is efficient.
- Other search algorithms can be integrated into KHM to achieve more efficient and effective clustering.
- XML Document can be processed with former Decompression.

Disadvantages of GSAKHM:

- GSAKHM requires more runtime compared to KHM.
- GSAKHM is not applicable when runtime is perilous.
- To store the labels of the nodes of document, storage space required is drastically increases as tree grows.

Ant based clustering was proposed which is based on the principal of ants behaviour and K-Harmonic Means algorithm (ACAKHM). It is based on the principle of collecting or organizing the feedback of the behaviour of the ants. Main advantage of ACA techniques is without defining the initial cluster centers, AC-Algorithm provides an appropriate partition of data. Grid size depends on number of objects and agent ants perform random walks on a grid on which the objects are dispersed randomly. Ants are allowed to move throughout the grid, selecting and dropping the objects inclined by the resemblance and density of the object. The likelihood of selecting up an object will be increased with low density neighborhoods, and decreased with high similarity among objects in the surrounding area.

The Ant clustering algorithm (ACA) possibly will take a long time to get a better result. To overcome drawbacks of ACA new algorithm is proposed which is based on Ant clustering algorithm and K-harmonic means clustering (ACAKHM), ACAKHM algorithm makes a better use of the advantage of both ACA and KHM; algorithm will escape from trapping in local optimal solution. In the interim, KHM algorithm can receive good initializations from the ACA, and provide better input to ACA [5].

Merits of ACAKHM:

- Overwhelms the initialization sensitivity of KM and KHM.
- Extents to global optimal effectively.
- ACAKHM outperforms than KHM and ACA when data sets are normalized.

Demerits of ACAKHM:

- Runtime is more as compared to KHM and ACA.

Candidate group search is based on some selection rules to isolate the candidate groups for each center (CGSKHM), 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.

CGS as it overcomes from drawbacks that KHMs solution is easily fallen into local optimum. The Candidate Group Search algorithm uses K-Harmonic Means algorithm which primary algorithm to obtain centers is used to obtain the center of each groups. KHM’s solutions where it is group of centers, CGS disturbs the solution, and replaces the centroids, to escape from local optimum this is the main feature of CGSKHM. Considering 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 and find each candidate group for current centroid. Use these groups to select possible to change current center. When we checking through data set, if it is satisfied with the selective rules, it shall be characterized into candidate group and used to replace the relative center. CGS replaces one relative center at a time, and takes the new center as an initial solution [6].

Merits of CGSKHM:

- CGS algorithm performance is better with less computational time in clustering, specifically for bulky datasets.
- CGS can moderate the computational time since the searching size is smaller comparing with VNS.

Demerits of CGSKHM:

- CGSKHM does not pledge to overcome from local optima.

K-harmonic means and Particle Swarm Optimization (PSOKHM) algorithm is the combination of Particle Swarm Optimization and K-Harmonic Means algorithm PSO method is population based global optimization technique. Each element is a discrete, and the swarm is composed of elements. In PSO, the solution space of the problem is articulated as a search space. Each position in the search space is an interrelated solution of the problem. Particles cooperate to find the best position in the solution space. Depend on velocity each particle changes according to its velocity calculating velocity and position of particle [7].

Due to lesser numbers function estimations KHM algorithm tends to converge faster than the PSO, but main drawback of KHM is it gets stuck in local optima. Here hybrid clustering algorithm called PSOKHM, which maintains the qualities of KHM and PSO. Objective function of KHM is the fitness function of PSOKHM; KHM is applied for four iterations to the particles in the swarm where it is collection of particles. To improve fitness value Particle Swarm Optimization algorithm is applied for eight generations [8].

Merits of PSOKHM:

- Speed of computation is more.
- Using F-Measure parameter it out performs then PSO which leads to overcome from trapping into local convergence.

Demerits of PSOKHM:

- PSO Algorithm requires more run time to overcome from local optima.
- PSOKHM is not applicable for run time critical applications.

Variable neighborhood search for harmonic means clustering (VNSKHM) was proposed [9]. Variable neighborhood search is empirical to solve the problem of KHM clustering which is easily trap in local optima. Variable neighborhood search (VNS) is a meta-heuristic intended for resolving combinatorial and global optimization problems. The elementary idea is to continue to a systematic change of neighborhood within a local search algorithm. Sets of neighborhoods are usually prompted from metric function which is specified in solution space. The algorithm concentrate on search around the similar solution until another solution improved than the mandatory is found and then restart the search, or jumps there.

Merits of VNSKHM:

- VNS is heuristic approach based solution for solving the issues of KHM clustering.

Demerits of VNSKHM:

- Computation required is more.

3. PROPOSED WORK

The proposed work is based on utilizing the best features of K-Harmonic mean and Gravitational Search algorithm to optimize the results and genetic algorithm will be used to find accurate results of data clustering.

In K-Harmonic Mean algorithm, initially, the cluster centers C (i.e. randomly Select center) are taken then calculate the K-Harmonic mean (KHM) of data points, membership functions and weight function associated with data points. Depending on membership function, weights for each center re-calculate its location. Using Genetic Algorithm, refine the cluster results by taking results from one population to new population. Selection of new population is based on their fitness function for fast and efficient retrieval.

Using the Hybrid K-Harmonic Mean and gravitational algorithm with Genetic Algorithm, a hybrid approach for clustering is proposed, the KHM requires fewer function evaluations, and this algorithm tends to converge faster than the GSA algorithm, but usually it gets stuck in local optima. On the other hand, the GSA algorithm always converges to better global optima for its excellent global searching ability.

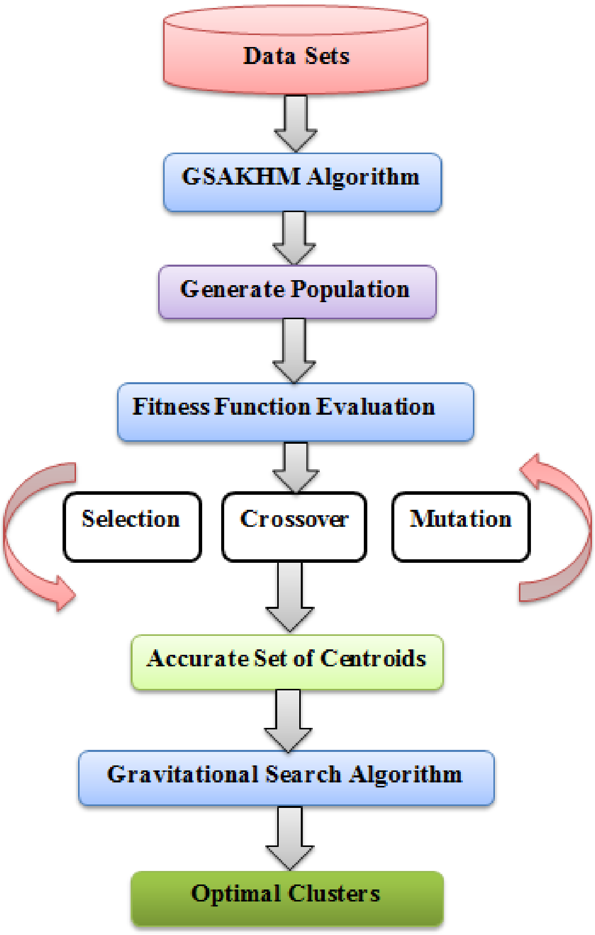

Figure 2: Architecture of HGSAKHM

Genetic Algorithm is adaptive, vigorous, effective, and global search methods for enhancing global search ability. They optimize a fitness function, corresponding to the preference criterion of data mining. To arrive at an optimal solution, uses certain genetic operator's like selection, crossover, and mutation. The complete flow of hybrid K-Harmonic Mean and gravitational algorithm for Evolutionary Algorithm is shown in See figure 2. New population is generated based on fitness evaluation [10].Using Genetic Algorithm optimal clusters can be obtained.

Advantages:

- Using the Hybrid K-Harmonic means and Evolutionary approach for Clustering algorithm leads to global optimal solution.
- By applying Genetic Algorithm accurate clusters can be formed.
- Optimal clusters can be obtained.

4. COMPARITVE STUDY OF KHM BASED CLUSTERING METHODS

The comparative analysis of various hybrid models of KHM is as shown in Table 1.Various parameters are considered for the comparison. The algorithms leads to run in local/global optimal solutions for searching where run in parameter is considered and KHM based clustering algorithms whether applicable to large datasets with their computation time required to run algorithm.

Table 1. Comparative study

Characteristics		Run in	Appropriate to large Datasets	Computation time Required
K-Harmonic Means based Clustering Approaches	**KHM K Harmonic Mean Algorithm**	Local optima	Not applicable for large data sets	More
	Gravitational Search Algorithm on KHM	Global optima	Applicable to large datasets	More but superior then PSOKHM and KHM
	Partial swarm optimization on KHM	Global optima	not applicable for large data sets	Requires More
	Candidate Group Search for KHM	Doesn't Guarantee Global optima	applicable for large data sets	Less Required
	Ant Clustering for KHM	Global optima	applicable for normalized data	Requires more time than KHM
	Variable neighborhood search for KHM	Global optima	Applicable for large Datasets	More

5. CONCLUSSION

In this paper several K-Harmonic Means based Clustering Approaches are discussed, with their merits and demerits. The various hybrid clustering techniques like GSAKHM, ACAKHM, PSOKHM method which performs clustering efficiently, used K-Harmonic Means as the objective function. KHM is independent of clusters center initialization and using KHM we can formulates effective and efficient clusters. Proposed model focuses on hybridization of KHM, GSA and GA for global optimization in clustering. The future extension may be to integrate KHM with other local search algorithms. Further, In future, the proposed hybrid method can be modified further for fast processing of large scale datasets using parallel architecture.

6. REFERENCES

[1] International journal of emerging technology and advanced engineering ‘comparison of various clustering algorithm of weka tools’ may 2012.

[2] Bangorn klabbankoh, ouen pinngern ph.d. An Applied Genetic Algorithms in Information Retrieval, king mongkut’s institute of technology ladkrabang ladkrabang bangkok 10520.

[3] International journal of emerging technology and advanced engineering “comparison of various clustering algorithm of weka tools” may 2012.

[4] Bin Zhang, Meichun Hsu, K-Harmonic Means - A Data Clustering Algorithm Umeshwar Dayal Software Technology Laboratory HP Laboratories Palo Alto HPL-1999-124 October, 1999.

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

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

[7] Cui, X., and Potok, T. E. (2005), Document clustering using particle swarm optimization. In: IEEE swarm intelligence symposium. Pasadena, California. Dalli, A, Adaptation of the F-measure to cluster-based Lexicon quality evaluation, In EACL 2003, Budapest.

[8] Fengqin Yang a,b,*, Tieli Sun a, Changhai Zhang “An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization”,2009.

[9] Abdulrahman Alguwaizani, Pierre Hansen, Nenad Mladenovic, Eric Ngai “Variable neighborhood search for harmonic means clustering”, 2011.

[10] Habiba Drias, Ilyes Khennak, Anis B, “A Hybrid Genetic Algorithm for large scale Information Retrieval”, 978-1-4244-4738-1/09/ ©2009 IEEE.