New Genetic Gravitational Search Approach for Data Clustering using K-Harmonic Means

Anuradha D. Thakare
Computer Engg Department,
Pimpri Chinchawad College of Engg, Pune

C. A. Dhote
Dept. of Computer Science and Engg,
Prof. Ram Meghe Institute of Tech and Research Badner Amravati, India.

Rohini S Hanchate
Computer Engg Department,
Pimpri Chinchawad College of Engg, Pune

ABSTRACT
In this article the new hybrid data clustering approach, Gravitational Genetic KHM, based on Genetic algorithm (GA), Gravitational Search Algorithm (GSA) and K-harmonic Means (KHM) is proposed. Data Clustering is used to group similar set of objects into set of disjoint classes, object in class are highly similar than the objects in other classes. Among various clustering methods, KHM is one of the most popular clustering techniques. KHM is applied widely and works well in many fields, but this method runs in local optima.

In the proposed approach the merits of Genetic Algorithm are used to escape the KHM clustering from local optima and to overcome the slow convergence speed of GSA. This paper is presented as work-in-progress in which the work model is proposed and some intermediate results are discussed which in turn will be compared with existing hybrid algorithms. The results are tested on several datasets.

Keywords
K-Harmonic Means (KHM), Clustering, Gravitational Search Algorithm (GSA), Genetic Algorithm (GA).

1. INTRODUCTION
Clustering is grouping data points into set of classes so that the data points in each class are higher degree of similarity, yet dissimilar from the others. Clustering techniques are applied in many applications such as information retrieval, pattern matching, data mining etc. There are many techniques for data clustering are developed; K-Mean algorithm is one of the most commonly used algorithms which applicable to large amount of data. As K-Mean algorithm is having several drawbacks, which are overcome by using K-Harmonic Mean algorithm which is center based approach which partitions the data objects into k clusters However, both KM and KHM easily trapped into local optima to resolve this problem, some hybrid clustering algorithms have been introduced like Genetic K-Mean algorithm, PSOKHM, ACAKHM and CGSKHM to help KHM escape from local optima, results in better clustering [1].

In this paper, hybridization of Genetic algorithm with KHM to generate new population in which KHM function as fitness function to overcome the drawback of KHM algorithm and GSA optimization algorithm which is based on laws of gravity, In GSA each individuals are having certain mass and force of attraction interact with each other, in this case we explore GSA and GA algorithm to achieve optimal cluster centers GSA is applied and to find optimal clusters Genetic operators are introduced to improve the GSA algorithm [2], called Genetic GSAKHM.in this way, a new hybrid data clustering algorithm based on KHM, genetic algorithm and GSA, is proposed and shown to be efficient. The experimental results are obtained by testing several data sets, which indicates Genetic GSAKHM superior to KHM.

Our remaining paper is further structured in this way. In section II, K-Harmonic Mean clustering method is discussed; hybrid clustering is proposed in section III, experimental results are discussed in section IV and conclusion about the paper is made in section V.

Our goal is to achieve fast, accurate and efficient data clustering by hybridization of K-Harmonic Means Gravitational Search Algorithm with Genetic Algorithms.

2. K-HARMONIC MEAN CLUSTERING
KM is partition algorithm which is simple commonly used clustering algorithm due to its implementation and small number of iterations, KM algorithm performance is based on the initialization of the centers which is major issue.

The K-harmonic mean algorithm [3] address this intrinsic problem by replacing the minimum distance by harmonic mean of distance from each point to all centers .if data point is closer to any one center harmonic mean gives a low score which is main property of KHM which is independent of center initializations.

Algorithm: K-Harmonic Means Algorithm[4]

Steps:

1. Randomly choose the initial centers.
2. Repeat
3. Calculate objective function value according to KHM function.
4. For each data points,
   a. compute its membership m in each center cj
   b. data points xi, compute its weight w(xi)
5. For each center cj, according to their memberships and weights re-compute its location from all data points xi:
6. Until KHM(X, C) does not change significantly.
7. Assign data point xi to cluster j with the biggest m membership.

3. PROPOSED HYBRID CLUSTERING APPROACH

The proposed work is hybridization of K-Harmonic Means and Genetic Algorithm to produce optimal set of centroids and further for cluster formation, Gravitational Search algorithm [5] is used which produces optimal clusters by removing the outliers. The algorithm for proposed work is as shown below:

Algorithm: Gravitational Genetic KHM

Steps:

1. Applying GSA.
   a. Force calculation for each data points.
   b. Calculate Force based on Euclidian distance between objects and mass of objects i and j, Total force acted on the dimension object i is calculated.
   c. Objects new position is calculated based on velocity at which they move.
2. Applying KHM objective function.
3. Using genetic algorithm finding optimal cluster centers.
   a. Final set of centroids.
   b. Calculate the total force acted on the dimension object i.
4. Applying GSA to get optimal clusters.

K-Harmonic Means Algorithm is independent of center initializations which calculate K-Harmonic Mean for each data points and depend on membership function data points assigned to centers. GSA is based on Gravitational law where each data points position and velocity is calculated when object moves there current position and new position is calculated. The object with heavier mass attracts the object with lower mass value; total force acted on each center is calculated. Genetic Algorithm is used to produce accurate cluster by using Selection Crossover and Mutation operators, KHM function act as fitness function to generate new population.

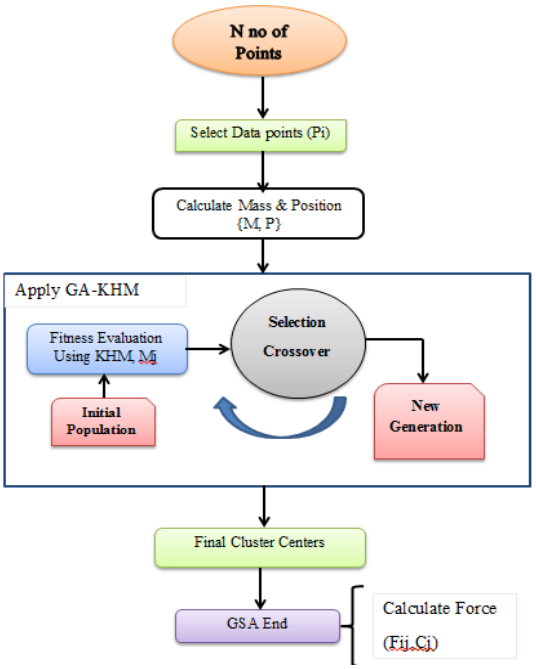

Fig 1: Clustering Model for Gravitational Genetic KHM

The graphical representation of working model of proposed approach is as given in fig 1. The work flow is divided into three parts:

Part A: Initially, the steps of GSA are executed i.e. for each data point in the population the mass and position is calculated by the following formulas [6].

$$Xi(t+1) \leftarrow Xi(t) + Vi(t) \quad \text{.....(1)}$$

Part B: Hybrid genetic KHM

In this, the GA steps are executed. The fitness values for each data point are calculated by KHM values [7].

$$KHM(X,C) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}} \text{.....(2)}$$

All the fitness values are now ranked to decide the highly fit members. On less fit members the GA operator's, selection, crossover and mutation are applied till termination criteria. The final cluster centers are obtained by this hybrid approach.

Part C: At the end, again we switch to GSA steps.

For all the final centroids calculate the acceleration, force and total force by the following formulas [8]:

$$a_i^d(t) = \frac{F_i(t)}{M_i(t)} \text{.....(3)}$$

$$F_{ij}^d(t) = G(t)\left(\frac{(M_i(t) * M_j(t))}{R_{ij}(t)+\varepsilon}\right)\left(X_j^d(t) - X_j^d(t)\right) \text{.....(4)}$$

And

$$Fi^d(t) = \sum_{j=1, j\neq i}^{N} rand_j \; F_{ij}^d(t) \text{.....(5)}$$

Finally, we get the optimal and accurate clusters which will be examined by the performance metrics like f-measure, KHM values and runtime factor.

4. EXPERIMENTAL RESULTS

The proposed algorithm is tested on six data sets and compared with other KHM based algorithms. These data sets are Iris, Breast Cancer, liver disorder, lung cancer, Wine and Glass. The details of these data sets can be viewed in table 1, the experimental results of simulation in table 2 and table 3 for different input parameters.

Table 1: Dataset description

Sr No	Name of dataset	No of classes	No. of features	Size of data set
1	**Iris**	3	4	150(50,50,50)
2	**Glass**	6	9	214(70, 17,76, 13, 9,29)
3	**Wine**	3	13	178 (59, 71,48)
4	**Breast cancer**	2	10	286(201,85)
5	**Lung cancer**	3	57	32(9,13,10)
6	**Liver disorder**	2	7	345(147,198)

4.1 Data Sets: Six real time datasets are taken for the experimentation [9, 10].

1. Iris data set consists of (n = 150, d = 4,k = 3), three different classes are Iris Setosa, Iris Versicolour and Iris Virginica. Each class contain 50 data points with four features.
2. Wine data set consist of (n = 178,d = 13,k = 3) three different classes class1 59 data points, class2 71 data points and class 3 contain 48 data points. With 13 features.
3. Glass data set consist of (n = 214,d = 9,k = 6) six different classes building windows float processed (70 data points), building windows non-float processed (76 data points), vehicle windows float processed (17 data points), containers (13 data points), tableware (9 data points), and headlamps (29 data points) with nine features.
4. Breast cancer (n=286, d=10, k=2) two different classes with 201 data points in one class 85 data points in other class with 10 features.
5. Lung cancer data set consist of (n=32, d=57, k=3) three classes with class1 9 data points class2 contain 13 data points and class3 contains 10 data points with 57 features.
6. Liver Disorder data set consist of (n=345, d=7, k=2) two classes class1 147 data points class2 contain 198 data points with 7 features.

4.2 Performance measures

The performance of proposed approach is measured by following metrics.

- K-Harmonic Mean (KHM) parameter is harmonic average distance from all data points to cluster centers.it clearly indicates that lower the KHM values Higher the quality of cluster is.
- F-measure uses the concept of precision and recall from information retrieval the F-measure is given by

$$F = \sum_{x} Max_{y}\{F(x,y)\}$$

Higher is F-measure higher is the quality of the cluster.

- Runtime parameter is used to measure the total time required to cluster the data sets.

4.3 Discussion on results

The intermediate clustering results are calculated with the input parameter P=2 and P=4. As discussed in section 3 this work includes the implementation of three algorithms, KHM, GSA and GA. Initially, KHM and Gravitational KHM are implemented as work-in-progress model and the results are finalized as intermediate results. These results are tabulated in table 2 and table 3 and will be further used for applying GA. The results shows that the KHM values and runtime required for KHM algorithm are less than the Gravitational KHM for almost all the data sets but, it runs in local optima therefore the quality of clustering affects. To overcome this Gravitational KHM is applied. The F-measure values for are Gravitational KHM is better than KHM algorithm which reflects accurate clusters except liver disorder dataset. The runtime parameter is considered as a result for both the algorithms and graphically represented in fig. 2 and fig. 3. It is observed that the runtime values for Gravitational KHM are more than KHM but it produces accurate clustering results.

Table 2: Results of proposed method with input parameter, p=2

Datasets	Criteria	KHM	Gravitational KHM
Iris	**KHM (X,C)** **F-measure** **Runtime**	**1.940** **0.863** **1.232**	**700.64** **0.920** **2.714**
Glass	**KHM (X,C)** **F-measure** **Runtime**	**0.012** **0.949** **3.463**	**702.361** **0.9897** **3.994**
Wine	**KHM (X,C)** **F-measure** **Runtime**	**0.045** **0.869** **1.856**	**704.066** **0.9440** **4.461**
Liver Disorder	**KHM (X,C)** **F-measure** **Runtime**	**40.92** **0.949** **2.402**	**703.096** **0.7884** **2.402**
Lung cancer	**KHM (X,C)** **F-measure** **Runtime**	**0.002** **0.343** **1.138**	**703.0933** **0.972** **0.187**
Breast cancer	**KHM (X,C)** **F-measure** **Runtime**	**0.014** **0.408** **3.775**	**702.809** **1.00** **5.444**

Table 3: comparison of various algorithm based on P where p=4

Datasets	Criteria	KHM	Gravitational KHM
Iris	**KHM (X,C)** **F-measure** **Runtime**	**117.880** **0.9493** **0.078**	**704.482** **0.8471** **2.799**
Glass	**KHM (X,C)** **F-measure** **Runtime**	**0.0110** **0.869** **4.509**	**704.961** **0.8409** **3.822**
Wine	**KHM (X,C)** **F-measure** **Runtime**	**0.0192** **0.846** **3.560**	**700.757** **0.9095** **4.956**
Liver Disorder	**KHM (X,C)** **F-measure** **Runtime**	**41.8609** **0.8493** **0.063**	**703.583** **0.6421** **0.546**
Lung cancer	**KHM (X,C)** **F-measure** **Runtime**	**0.0081** **0.544** **1.908**	**701.037** **0.6824** **2.352**
Breast cancer	**KHM (X,C)** **F-measure** **Runtime**	**0.1145** **0.523** **4.107**	**702.606** **0.839** **5.304**

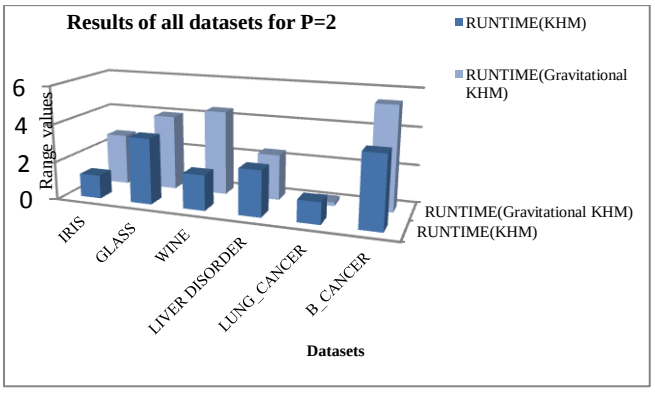

Fig 2: Results datasets for input parameter P= 2

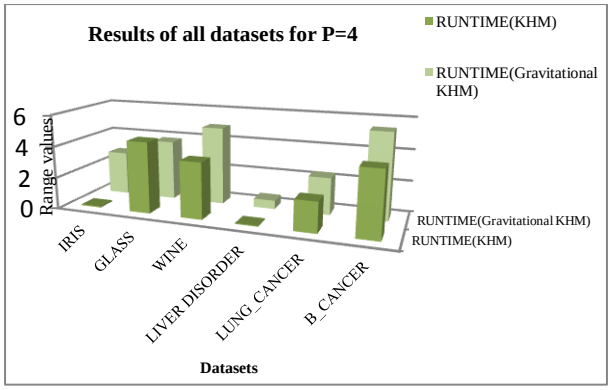

Fig 3: Results datasets for input parameter P= 4

The intermediate results are calculated for KHM and Gravitational KHM as shown in table 2 for input parameter p=4 and in table 3 for p=3. Quality of cluster is evaluated with KHM, F-measure and Runtime parameters. In both the tables good values for intermediate results are obtained. Therefore the best results are expected after applying GA. These results then will be compared with the existing hybrid algorithms.

5. CONCLUSION

In this paper hybrid Gravitational Genetic KHM clustering method based on KHM and GSA algorithm is proposed. The proposed algorithm is tested on six data sets for intermediate parts and experimental results show that the algorithm gives both efficient and accurate clustering.

The experimental results shows that the Gravitational KHM produces accurate clustering results in terms of F-Measure, though the runtime required is more than KHM. This will be further improved by our proposed method using Genetic Algorithm.

It is expected that Proposed Gravitational Genetic KHM will definitely results in good quality clusters.

6. REFERENCES

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

[2] International journal of emerging technology and advanced engineering "comparison of various clustering algorithm of weka tools" may 2012.

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

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

[5] Minghao Yin, Yanmei Hu, Fengqin Yang, Xiangtao Li, Wenxiang Gu A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering College of Computer Science, Northeast Normal University, Changchun 130117, China ,2011.

[6] An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization Fengqin Yang , Tieli Sun, Changhai Zhang 2009.

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

[8] On The Performance of the Gravitational Search Algorithm, Taisir Eldos Department of Computer engineering College of computer engineering and Sciences ,Rose Al Qasim IJACSA Vol. 4, No. 8, 2013.

[9] Data sets from http://archive.ics.uci.edu/ml/datasets.

[10] 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.