Journal of Theoretical and Applied Information Technology
10th June 2015. Vol.76. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

META PHYSICAL ALGORITHMIC REPRESENTATION FOR FLAWLESS CLUSTERING

K. MOHANA PRASAD[1], DR. R. SABITHA[2]

[1]Department of CSE,Sathyabama University, Chennai 600119,Tamil nadu,india
Head and professor, Dept of IT, Jeppiaar Engineering College,Chennai 119. Tamil nadu,india
E-mail: mohanaprasad1983@gmail.com *, sabitha-ramadoss@yahoo.com[2]

ABSTRACT

Data mining is defined as analyzing data from a large database .The task performed by data mining is clustering, which are used for grouping the data objects to form a new class .The two algorithms of clustering is MH-means (modified harmonic means) and GSA(gravitational search algorithm).MH-means is the algorithm with some random based centroid for desired clusters. In this algorithm every data object are assigned to the closest centroid. It is used to calculate the mean value of data objects within the clusters , so that they reach an optimal solution. MH-meansis an iterative process. GSA is based on interaction of masses in universe through Newtonian Gravity law. The role of GSA is to identify the problem in space. By combining the MH-means and GSA algorithms the performance is increased. Along with this combination a new method Standard Dataset is also used. The standard dataset are used for increasing the convergence speed of the proposed methods.

Keywords: *Clusters ,Centroid, Gravity Law ,GSA , MH-means.*

1. INTRODUCTION:

Nowadays data extracted from various or large database are transformed into meaningful structure. This transformed structure is used for various purpose and its powerful and produces an intended result. The size and difficulty of performances in the datasets are increased. The KDD(knowledge discovery dataset) are cleaning the missing values ,inconsistent and incomplete data, integrating multiple values ,selecting the relevant data, transforming to suitable format, knowledge extraction from intelligence ,some interesting measures or thresholds are applied and exact pattern returned ,presentation in graph trees etc.. The above KDD are represented in many ways. More functionality is also used.

Clustering is used by many applications. It is said to be an attractive task in data mining. The major uses of clustering are marketing ,land use ,insurance, city planning and earth –quake studies. Data are grouped to form new classes and also form a set of meaningful subclasses. The clustering principles are higher in intra-classes and lower in inter-classes. Clustering is said to be an unsupervised classification because it recognizes the pattern without guideline information. The three main functions used in clusters are i) Initialization, ii)Choosing random datasets ,iii)Comparison.

Clustering has various algorithms. A larger variety of applications are developed with the help of these algorithms. Partitioning and Hierarchical are the algorithms that are used . Partitioning algorithms are H-means ,where the mean value of the objects represents each cluster center. H-mediods is where the data points belongs to the mediods, if the random objects are assigned to any of the mediods then the data point is reassigned to other mediods. CLARA (cluster large application) is used to select best

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

H-mediods among the selected sample of the dataset. CLARANS(clustering large applications based on randomized search) draws a sampke at any random order at every step of the search and then applies PAM(partitioning around mediods) which has fixed samples at every stage of the search. Hierarchical algorithm are top-down or bottom-up approaches ,another name for this approach is Agglomerative Nesting and Divisive Analysis. Minimum and maximum distance ,mean distance and average distance are calculated here. BIRCH(balanced iterative reducing and clustering using hierarchies) is based upon the clustering feature and clustering feature tree.

1.1 Related Articles:

In H-means clustering ,Lloyd's algorithm is used. A minor implementation of this algorithm is called Filtering algorithm. Using the filtering algorithm , its easy to obtain the Hd-tree as the only major data structure. The efficiency of this algorithm are related in two ways: Data sensitive analysis is used in filtering the running time. This shows about the increasing in the separation of clusters. From the applications in color quantization, data compression and image segmentation ,presents a number of empherical studies both on synthetically generated data and on real datasets. They donot pass information from one stage to the next. Centroids get converged to their final positions, so that they could contain the same data points when they converge. Hence the running time will be more in this algorithm. In this MH-means clustering is based on initial centroids that are selected randomly. To avoid this random selection of centroid , enhanced MH-clustering is used. This reduces the iteration and are lesser in time . In this ACO technique two phases are used i) statistical modes ii) refinement algorithm. The distance measure is used in max-min measure. The last iteration is the fixed size of the clusters. Harmonic H-means is an extension of standard MH-means clustering algorithm that is used to identify the non-linearly separable clusters. To overcome cluster initialization problem in this method, global harmonic MH-means clustering is used. The deterministic and incremental approach to Harmonic based clustering is used. One cluster is added at every stage of executions of Harmonic MH-means through global search procedure. They do not depend upon the initialization of clusters, it identifies non-linearly separable clusters and due to there incremental nature and search procedure they locate near optimal solutions by avoiding poor local maxima and also reduces the computational cost. Here elapsementto time is more. Since H-clustering is similar to spectral clustering, spectral methods is easy to weight the data points. Since some data points escape from clusters due to weak data points, optimal solutions are not fixed. An unsupervised algorithm is used for finding a hidden data from an unlabeled data. Clusters can be found easily by multidimensional data, by using the properties of Gaussian distribution from Gaussian model. The usage of different methods and formula helps to speedup the performance . By using image segmentation it compares against H-means and mean-shift algorithm. Many clusters are used in mean-shift algorithm that is used to increase time at last. The methods used in clustering are i) ACA ii) ABC iii) HABC iv) MHM v) GSA-MHM. Clustering is one of the technique or task of data mining , and also a collection of similar data.

2. EXISTING WORK:

In this they had five real datasets that was used and each datasets have different criteria: Best ,Average ,Worst ,STD,NFE (number of functions evaluation). When compared to other algorithms such as H-means ,Genetic algorithm, simulated annealing ,Ant Colony Optimization, HoneyBee Mating Optimization, and Particle swarm optimization, the combination of GSA-MHM performance level was higher and efficient. The functional evaluation is lesser in GSA-MHM.

Methods:

1) The total mean-square quantization error is used for goodness of clustering.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2) Euclidean distance formula is used for comparing and measuring the dissimilarity.
3) To calculate the fitness in criteria.

Let us consider set of n data objects as (T1,T2,….Tn). for each objects, described by d features and in matrix form (Tn x d). And each cluster c={c1,c2,c3….cp} for p groups. After assumption the summation of all clusters should be equal to the number of objects in the data set. The objects are represented in row and column.

For partitioning the dataset in different ways, fitness function is defined for quality purpose and goodness of cluster. MSE is defined as :

$$f(T,C)=\sum_{i=1}^{k}\sum_{T_i \in c_i} d(T_i, Z_i)^2 \quad (1)$$

where $d(T_i, Z_i)$ is the dissimilarity between object T_i and centroid of cluster Z_i by calculating mean value of objects within the cluster. And to calculate the dissimilarity between data objects for performances distance measure calculation is used.

Consider two objects as S_i and S_j with d dimensions, the second equation is defined as:

$$d(S_i,S_j)=\sqrt{\sum_{p=1}^{d}(s_i^p - s_j^p)^2} \quad (2)$$

These datasets like Iris, Wine, Glass, Cancer, CMC (Contraceptive Method Choice) were used for evaluating performance of clustering algorithms and other datasets.

3. PROPOSED SYSTEM:

In clustering, the principle maximizes the intra class and minimizes the inter class. By using the algorithm GSA-HM-MHM, in proposed system different data sets which is used for reducing iterations , which is present in UCI repository that has higher instants to reduce the iterations for using different method.

ALGORITHM: GSA-HM-MHM

Formation of data objects :

1)Consider the data objects be S={x1,x2,x3….xn}.Each data objects has ‘V’ features.

2)It is represented by the data matrix S(txv).For each row, d dimension is present. And for t row d dimension is used.

Clustering Algorithm :

3) To group the data object which is similar and dissimilar data objects.

Conditions :

3a)In the cluster, there should be at least one data objects.

3b)No data objects should be same in different clusters.

3c)After assigning the data objects to the clusters, the sum of all the objects should be equal to original dataset.

3c1)So that it is easy to partition the given datasets into many ways if it satisfies the three conditions.

Partitioning :

4)To measure the quality of cluster fitness function must be defined.

5)The famous function used is MSE(Mean Square Quantization Error) Eg: $d(t_i,z_l)$ to calculate the mean value for dissimilarity between objects and Centroid.

6)To calculate the dissimilarity between the objects , Harmonic mean Distance measure is used.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

H-Means:

1)After partitioning the cluster k-means is used. It starts with random Selection of centroid for desired cluster and assign every data objects which is closest to the centroid.

2)Iteration process takes place to reach the optimal solution.

GSA :

1) Gravitational Search Algorithm is used to search for the problem Space.

2) For finding GSA, need to find force ,acceleration, velocity and Weight(mass).

3) Force is calculated by masses and the distance between the objects. Acceleration depends on the force and the mass. So the objects is Heavier .

4) And also GSA is used to find the fitness value, for finding the maximum, average and minimum.

GSA condition :

1)Decreases number of iterations and function evaluation compared to original dataset.

2)A good initial population or cluster is needed so that it could easily search for near global optima solution.

GSA-HM-MHM

1)On selected dataset k-means algorithm is applied to find the centroid for the clusters.

2) Then Apply the GSA algorithm by using the normalization and Harmonic mean measures then calculate the classification rate according to the fitness value.

2a)Based on the dataset minimum, maximum, average values are Calculated.

2b) The other solution will be created randomly.

3)GSA will be used for optimal solution.

4)Then generate the levels used for classifying the clusters.

5)Convergence curve which shows the decreasing and increasing the Cluster formation.

Clustering Functions are used to improve the efficiency three steps are used:

1)Avoiding the initialization of k-value. foreg: if k=2,it should not assigned only for the given k-value in the data set.

2) Avoiding randomly chosen dataset so that iterations will be lesser.

3) Comparisonie, for example if we have 20 datasets, by choosing randomly the datasets say 1st,12th it should be compared separately .while comparing it the least value should be taken and again the iteration should starts.

The distance measure for the formation of clustering is : Harmonic formula :

$$HM=S/(\frac{1}{x1}+\frac{1}{x2}+\frac{1}{x3}+\frac{1}{x4}+\cdots+\frac{1}{xS})$$

Here, S is the total number of samples, x1, x2, x3… xN are the individual values. Number of objects will be averaged and divided by total number of values. To increase the convergence speed and reduction of iteration a new algorithm GSA-HM-MHM is used.

The values which calculated is in right form mean square error is calculated accordingly classification rate is calculated.

Journal of Theoretical and Applied Information Technology
10th June 2015. Vol.76. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. EXPERIMENTAL RESULTS:

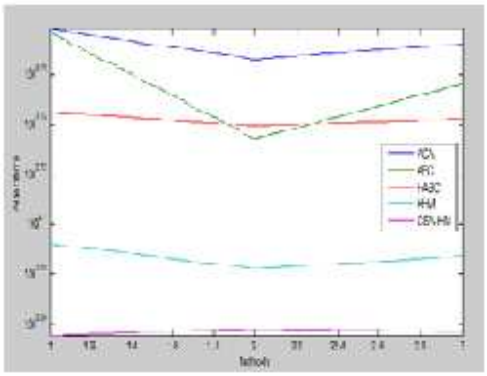

Fig 1-comparsion of different method and algorithm

Table 1- The performance of different values using objective function

Method	ACA	ABC	HABC	KHM	GSA-KHM
BEST	6000	5984	5552	4921	4523
WORST	5834	5421	5490	4812	4551
MEAN	5917	5703	5521	4867	4537

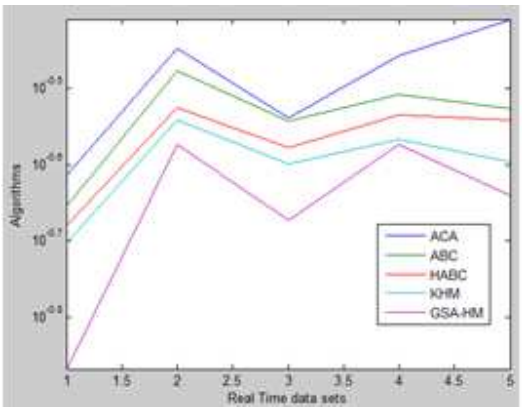

Fig 2-comparsion of different real time dataset using algorithm

Table 1- The Performance Of Different Real Time Dataset Using Algorithm

Real Time data sets	ACA	ABC	HABC	KHM	GSA-KHM
Iris	0.2431	0.2216	0.2091	0.1986	0.1354
Wine	0.3567	0.3341	0.2987	0.2879	0.2671
Vowel	0.289	0.2867	0.2645	0.2521	0.2121
Art	0.349	0.311	0.2923	0.271	0.2669
Wood	0.39	0.2978	0.2876	0.2531	0.2289

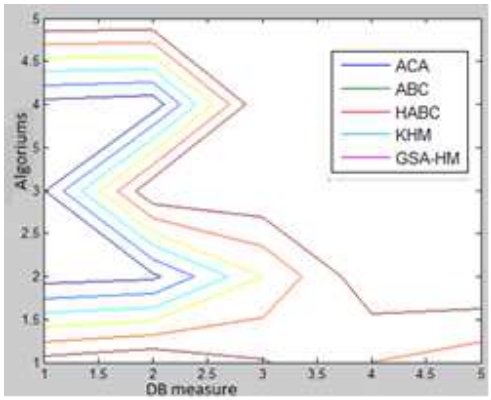

Fig 3-comparsion of different DB measurement using algorithm

Table 2- The Performance Of Different DB Measurement Using Algorithm

DB Measure	ACA	ABC	HABC	KHM	GSA-KHM
Iris	2.944	2.9979	2.9097	2.799	2.7354
Wine	2.349	2.377	2.6993	2.977	2.9999
Vowel	2.39	2.9978	2.9879	2.954	2.9989
Art	2.36	2.3347	2.9987	2.988	2.9977
Wood	2.989	2.9897	2.9645	2.96	2.9797

5. CONCLUSION

In this the clustering objects uses the hybrid combination algorithm(GSA-HM-MHM). The advantageous method to escape from the local optima is the H-mean method. This gives an optimal solution by reducing the iterations. The performance for different datasets are calculated . The various performances of datasets are used for research purposes. Functional values are calculated by the usage of Harmonic mean formula and the values are reduced when compared to the Euclidean distance measures.It is used for various purposes like image segmentation, bio-medical, Military Appliances.

REFERENCES

[1] K. Mohana Prasad, Dr. R. Sabitha "Evolution Of An Algoritm For Formulating Efficient Clusters To Eliminate Limitations" International Journal of Applied Engineering Research (IJAER), ISSN 0973 – 4562, Volume 9, Issue 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[2] K. Mohana Prasad , Dr. R. Sabitha (2014), 'Formation of Clusters through Analysing the Similarities and Strength of the Variables in a Data Set', International Conference on Circuit, Power and Computing Technologies [ICCPCT], ISBN: 978-1-4799-2395-3/14/$31.00 ©2014 IEEE pp.1728-1732.

[3] Ahmadi, F. Karray, M.S. Kamel," Model order selection for multiple cooperative swarms clustering using stability analysis", Information Sciences 182 (2012) 169–183.

[4] Akay, D. Karaboga, "A modified artificial bee colony algorithm for real-parameter optimization", Information Sciences 192 (2012) 120–142

[5] G. Babu, M. Murty, "A near-optimal initial seed value selection for k-means algorithm using genetic algorithm", Pattern Recognition Letters 14 (1993) 763–769.

[6] Bagirov,"Modified global k-means algorithm for sum-of-squares clustering problem", Pattern Recognition 41 (2008) 3192–3199.

[7] Z. Che, A. Unler," Clustering and selecting suppliers based on simulated annealing algorithms", Computers and Mathematics with Applications 63 (1) (2012) 228–238

[8] J. Jimenez, J. Mares, V. Torra, "An evolutionary approach to enhance dataPrivacy", Soft Computing 15 (7) (2011) 1301–1311.

[9] T. Li, N. Li, J. Zhang, I. Molloy, "Slicing: a new approach for privacy preservingData publishing", IEEE Transactions on Knowledge and Data Engineering 24 (3)(2012) 561–574.

[10] N.V. Mogre, G. Agarwal, P. Patil, "A review on data optimization technique forData publishing", International Journal of Engineering Research & Technology 1(10) (2012).

[11] C.C.Hsu,Y.C.Chen, "Mining of mixed data with application to catalog marketing", ExpertSyst.Appl.32(1)(2007)12–27.

[12] W.Chen,G.Feng,"Spectralclustering:a semi-supervised approach", Neuro- computing 77(1)(2012)229–242. 11.

[13] P.Li,C.Chen,J.Bu,"Clustering analysis using manifold kernel concept factorization", Neurocomputing87(15)(2012)120–131.

[14] Ahmadi, F. Karray, M.S. Kamel, "Model order selection for multiple cooperative swarms clustering using stability analysis", Information Sciences 182, (2012) 169–183.

[15] Z. Che, A. Unler, "Clustering and selecting suppliers based on simulated annealing algorithms", Computers and Mathematics with Applications 63 (1) (2012) 228–238.

[16] H. He, Y.Tan, "A two-stage genetic algorithm for automatic clustering", Neurocomputing 81(1) (2012) 49–59.

[17] P.Li, C.Chen,J. Bu, "Clustering analysis using manifold kernel concept factorization, Neuro computing 87 (15) (2012) 120–131.