A HYBRID K-HARMONIC MEANS WITH ABCCLUSTERING ALGORITHM USING AN OPTIMAL K VALUE FOR HIGH PERFORMANCE CLUSTERING

Sithara E.P and K.A Abdul Nazeer

Department of Computer Science and Engineering, National Institute of Technology, Calicut, Kerala, INDIA

ABSTRACT

Large quantities of data are emerging every year and an accurate clustering algorithm is needed to derive information from these data. K-means clustering algorithm is popular and simple, but has many limitations like its sensitivity to initialization, provides local optimum solutions. K-harmonic means clustering is an improved variant of K-means which is insensitive to the initialization of centroids, but still in some cases it ends up with local optimum solutions. Clustering using Artificial Bee Colony (ABC) algorithm always gives global optimum solutions. In this paper a new hybrid clustering algorithm (KHM-ABC) is presented by combining both K-harmonic means and ABC algorithm to perform accurate clustering. Experimental results indicate that the performance of the proposed algorithm is superior to the available algorithms in terms of the quality of clusters.

KEYWORDS

Data Mining, Clustering, K-means Clustering, K-Harmonic means Clustering, Artificial Bee Colony Algorithm

1.INTRODUCTION

Cluster analysis is one of the important data analysis method which is used in the areas like data mining, vector quantization, image analysis and compression. Clustering is a process which sequentially takes data as inputs and outputs clusters as results. The aim of clustering is to assemble a set of similar objects into a group that in some sense belong together because of related characteristics [1][2].

Among the various clustering methods available, K-means is very simple clustering method to cluster data sets, but this method highly depends on the initial selection of centroids and usually converges to the local optimum solutions [1][3][4][5]. Similarly, K-harmonic Means clustering is a centroid based clustering algorithm in which the harmonic mean of the distances between the centroids and each data point is used as the main component of the performance function [4][6].

K-harmonic means provide better clustering results than K-means [2][3]. In K-means the bond between data points and the nearest centroid is very strong, so that data point is strongly attached to a cluster centroid and it is moving to another cluster only when it is too close to another centroid. This powerful bond stops the centroids from shifting out of the surrounding locality of data. In K-harmonic means, the harmonic means function is used to establish the link between data points and centroids. This association is distributed and make the algorithmnsensitive to

DOI: 10.5121/ijci.2016.5206

initialization. In [7], Bin Zhang provided a performance chart for sensitivity to initialization, given in Figure 1. In this paper KHM_p is the K-harmonic means clustering algorithm in which the p^{th} power of distance function is used, usually p = 2. The K-harmonic algorithm always converge faster than K-means, but sometimes it provides less accuracy clustering results. If the problem contain many local minima then the algorithm will fall into a local optimum solution [8].

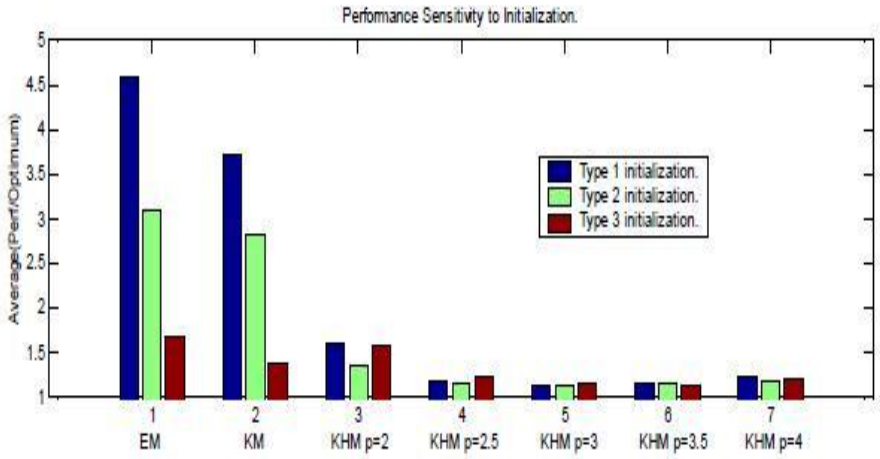

Figure 1. Performance sensitivity to initialization [7]

A better mode of clustering is by the use of Artificial bee colony algorithm (ABC). ABC was introduced by Dervis Karaboga in 2005 [9][10] which was established on the foraging characteristics of honey bees, but can be used for solving numerical optimization problems. It is a population based optimization algorithm. Here, the primary idea is to use ABC to generate best solution by providing non-local moves for the cluster cores

2. RELATED WORK

Several researchers have highlighted their work in the field of clustering. The continuing work on this area has brought about novel and enhanced methods for clustering.

2.1. K-HARMONIC MEANS CLUSTERING

K-harmonic means algorithm is a centre based clustering algorithm, in which the harmonic means between centroids and data values are taken as the main constituent of the performance function [6][7]. The performance function is given in equation 1, where x_1 to x_n are data points, c_1 to c_k are centroids and k is the required number of clusters.

$$perf_{KHM} = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{||x_i - c_j||^2}} \qquad (1)$$

It was demonstrated that K-harmonic means is essentially insensitive to the centroid initialization. In K-means, arithmetic mean is used and the value is always near to the higher of two values. But harmonic mean value is near to the minimum of two values. This property enables K-harmonic means clustering to suppress the effect of outliers. The performance of K-harmonic means in improving the quality of clusters, was better than K-means.

The hybrid clustering algorithm introduced by Ravindra Jain [9] is based on applying K-means and K-harmonic means in tandem to find cluster mean until termination condition. It shows that K-harmonic means yields better accuracy in the clustering than K-means algorithm.

Fangyan Nie, Tianyi Tu, et al. proposed a combined Particle swarm optimization and K-harmonic means clustering (PSOKHM) algorithm in [8]. It was a hybrid algorithm which combines the benefits of both algorithms. K-harmonic means clustering sometimes fall into local optimum solutions, so a Particle swarm optimization (PSO) was used to evolve non-local centroid movements and leads to better solutions. In this hybrid algorithm, two methods are used to find the cluster means, thus the algorithm produces better results because it combines the advantages of both the techniques.

2.2. ARTIFICIAL BEE COLONY ALGORITHM

Artificial bee colony (ABC) algorithm was introduced as a swarm-based algorithm [10]. The algorithm is explained by categorising the bees into three groups: the employees, onlookers and scouts. The number of employed bees and the number of onlooker bees are same as that of the number of solutions. In other words, the number of solutions are equal to the number of food sources around the hive.

Employed bees are searching for the food sources. After identifying a food source it returns back to the hive and dance in the dancing area of the hive. Onlooker bees observe this activity to get an idea about the nectar quantity of food sources and choose a food source with a good amount of nectar. The food source with least amount of nectar is considered as abandoned and that bee is acting as a scout and starts to search for a new food source. A possible solution to the problem is represented by the position of food source and the quality (fitness) of a solution is represented by nectar amount of the food source.

In [10] Changsheng Zhang, et al. discussed how ABC can be used for clustering and they proved that the ABC algorithm can work effectively, by analysing the computation time of the ABC algorithm and other well-known techniques [10][11]. A performance comparison table is given in Table 1 taken from [10]. It shows that the ABC algorithm achieves better results. In [12] Bahriye Akay and Dervis Karaboga compared ABC algorithm with other algorithms like genetic algorithm, evolutionary algorithm, particle swam optimization, etc. They proved that ABC algorithm has better performance compared to the mentioned algorithms.

Table 1. Performance comparison table [10]

The average fitness computation numbers and computation time.

Data set		GA	ACO	K–NM–PSO	ABC
Iris	Time (s)	105.53	33.72	48.13	**29.68**
	Numbers	38128	10998	**4556**	8658
Thyroid	Time (s)	153.24	102.15	118.46	**85.26**
	Numbers	45003	25626	**7245**	24136
Wine	Time (s)	226.68	68.29	589.40	**48.85**
	Numbers	33551	9306	46459	17554

Giuliano Armano and Mohammad Reza Farmani [3] proposed a hybrid algorithm as a

combination of artificial bee colony algorithm and K-means algorithm. K-means algorithm is highly dependent on the initialization of centroids and usually gets stuck in local optima. The ABC algorithm performs a global search in the entire solution space and it can generate good and global results. The authors propose a new combination algorithm which makes use of the combined benefits of K-means and ABC algorithms for solving clustering problems.

3. PROPOSED APPROACH

Even though K-harmonic means is insensitive to initial-ization of centroids, the cluster quality needs to be improved by finding global optimum solutions. Thus a well performed optimization algorithm, ABC algorithm is utilized for non-local movement of centroids and to obtain more promising results.

In this hybrid approach k value should be fixed before executing the algorithm. Gap statistics method and Average silhouette width method are used to identify the optimal k value and the value thus obtained is used to fix the number of initial food sources in the proposed algorithm.

3.1. IDENTIFYING OPTIMAL K VALUE

At the pre-processing stage optimal k value is estimated using Gap Statistics method [13]. The obtained value is verified using Average silhouette width method [14].

3.1.1. GAP STATISTICS METHOD

Run a K-means algorithm on the given set of data to find number of clusters, and sum the distance of all points from their cluster mean, this is the dispersion. Generate some number of sample data sets of original and find the mean dispersion of these sample data sets. Each gap is defined as the logarithmic difference between the mean dispersion of reference data sets and dispersion of the original data set. Take the minimum value of k for which the gap is maximized.

3.1.2. AVERAGE SILHOUETTE WIDTH METHOD

Run a PAM (Partition Around Medoids) algorithm on the original data set for values of k, in the range 2 to 10. The average silhouette width of clusters formed is calculated for each iteration. Observe the highest value. The k value corresponding to the highest average silhouette width is taken as the optimal k value.

3.2. PROPOSED ALGORITHM

In the proposed approach, ABC algorithm helps the K-harmonic means clustering algorithm to set the global optimum solutions. The K-harmonic means performance function is used to calculate the fitness of each solution and the characteristics of ABC algorithm leads to global optimum solutions rather than local optimum solutions. These properties of both the algorithms provide better performance in the quality of clusters. The method is formulated in algorithm 1

Algorithm 1 Pseudo-code of the proposed algorithm

Input : Number of data values indicated as $x_i..x_n$

: Values for control parameters SN (number of food sources which is same as k), limit and MCN (maximum cycle number).

Output: SN number of clusters.

1) Begin

2) Initialize trial counter array with values zero
3) Load data set values.
4) Initialize food sources(centroids) c_i where i = 1…SN

5) Evaluate the fitness values (fit_i) of the food sources using k-harmonic means performance function.

6) Set cycle to 1
7) Repeat until the termination criteria met (cycle = MCN)

8) For each employed bee

   a) Produce new food source (v_{ij}).
   b) Use k-harmonic means to evaluate the new fitness values.

   c) Compare them with the original one, if the fitness value does not improve increment the corresponding trial counter value.

   d) Better food source will be memorized and delivered to onlooker bee.

9) Evaluate probability values (p_i) of food sources.

10) For each onlooker bee

   a) Select a food source depending on probability.
   b) Produce new food source.
   c) Apply k-harmonic means to find new fitness values

   d) Compare them with the original one, if the fitness value does not improve increment corresponding trial counter value and memorize the best food source

11) Check if trial counter value > = limit value then the food source is abandoned by the bee and that employee bee become scout. Abandoned solution is replaced by the scout with new randomly produced food source.

12) Memorize the best solutions achieved so far
13) cycle=cycle+1
14) End

Fitness calculations are given in equations 2 and 3.

$$f_i = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{||x_i - c_j||^2}} \qquad (2)$$

$$fit_i = \frac{1}{1 + f_i} \qquad (3)$$

Probability equation and equation to produce a new random solution are given by equations 4 and 5 respectively ϕ_{ij} is a random number between -1 and 1.

$$p_i = \frac{fit_i}{\sum_{m=1}^{k} fit_m} \quad (4)$$

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \quad (5)$$

4. EXPERIMENTAL RESULTS

This section presents the outcome of the experiments carried out for evaluating the performance of the suggested algorithm in improving the cluster quality.

4.1. ALGORITHM IMPLEMENTATION

The standard K-harmonic function and the proposed algorithm (KHM-ABC) were coded in R programing. The data sets used are iris, wine, yeast and spam base downloaded from UCI learning repository.

4.2. RESULTS

4.2.1. OPTIMAL K VALUE IDENTIFICATION

Gap statistics method is used to identify the optimal k value. A plot for wine data set is shown in Figure 2. From the figure, it clearly shows that the minimum value of k with maximum gap is 3. Thus the optimal value of k is 3.

4.2.2. VERIFYING THE OPTIMAL K VALUE

The obtained k value is verified using Average silhouette width method and the plot is given in Figure 3 for wine data set. From the given plot we can infer that the highest average silhouette width value is found at k = 3. Thus the optimal k value can be taken as 3.

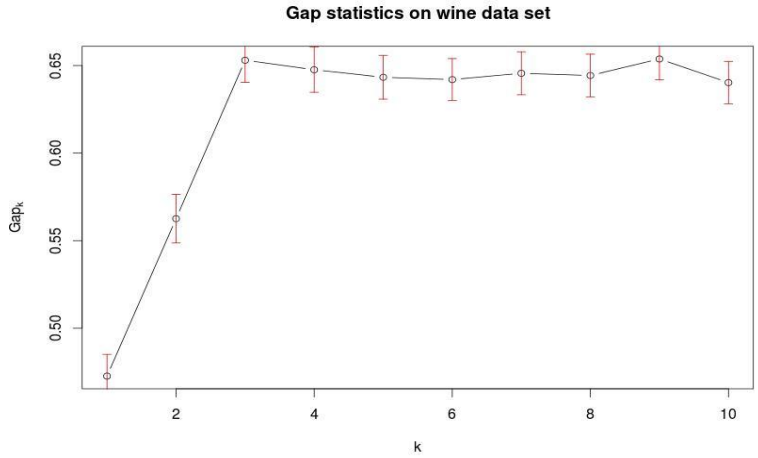

Figure 2. Gap statistics for wine data

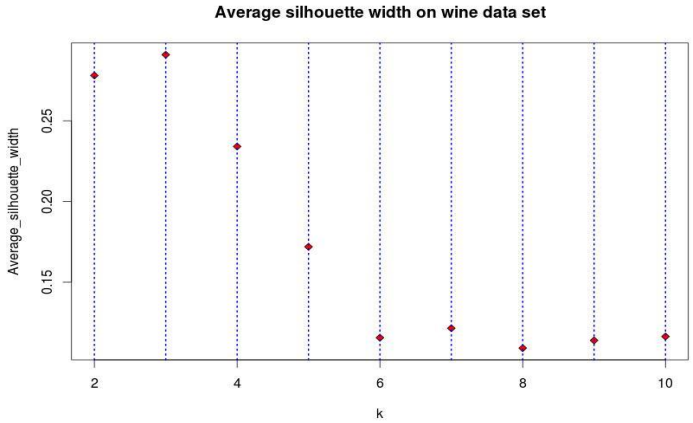

Figure 3. Average silhouette width for wine data

4.2.3. PERFORMANCE SCORES

Silhouette index scores are used to evaluate the performance. Silhouette index scores for clustering algorithms K-means, K-harmonic means, PAM, ABC and KHM-ABC are calculated on different data sets iris, wine, yeast and spam base. The results of the experiments are tabulated and given in Table 2.

The performance comparison graph is given in Figure 4, which shows the improvement of KHM-ABC algorithm in terms of accuracy.

5. CONCLUSIONS AND FUTURE SCOPE

K-harmonic means algorithm overcomes many of the limitations of K-means, but still it may get trapped into local optimum solutions. The proposed method (KHM-ABC) used artificial bee colony algorithm to optimize K-harmonic means clustering algorithm to improve the clustering quality. ABC algorithm always provides global optimum solutions. This feature helps K-harmonic algorithm to fix a good set of initial centroids. The proposed method guarantees the cluster quality. Cluster quality was checked using silhouette index scores. Silhouette index scores are calculated for KHM-ABC and other related popular algorithms ABC, K-means K-harmonic means and PAM. The results showed that the performance of KHM-ABC was better compared to

the other algorithms.

One of the main constraints of the proposed algorithm is that the value of k is not self-learned. In the pre-processing stage the k value was fixed using gap statistics method. The k value thus obtained is verified using silhouette width method. Some statistical method with a systematic approach is worth investigating for determining the value of k at run time.

Table 2. Performance comparison table

Sl.no.	Method	Data set	Optimal k	Silhouette index score
1	K-means	Iris	3	0.55
		Wine	3	0.57
		Yeast	6	0.16
		Spambase	3	0.68
2	KHM	Iris	3	0.55
		Wine	3	0.57
		Yeast	6	0.15
		Spambase	3	0.67
3	PAM	Iris	3	0.55
		Wine	3	0.57
		Yeast	6	0.15
		Spambase	3	0.68
4	ABC	Iris	3	0.53
		Wine	3	0.56
		Yeast	6	0.16
		Spambase	3	0.66
5	KHM-ABC	Iris	3	0.55
		Wine	3	0.57
		Yeast	6	0.2
		Spambase	3	0.71

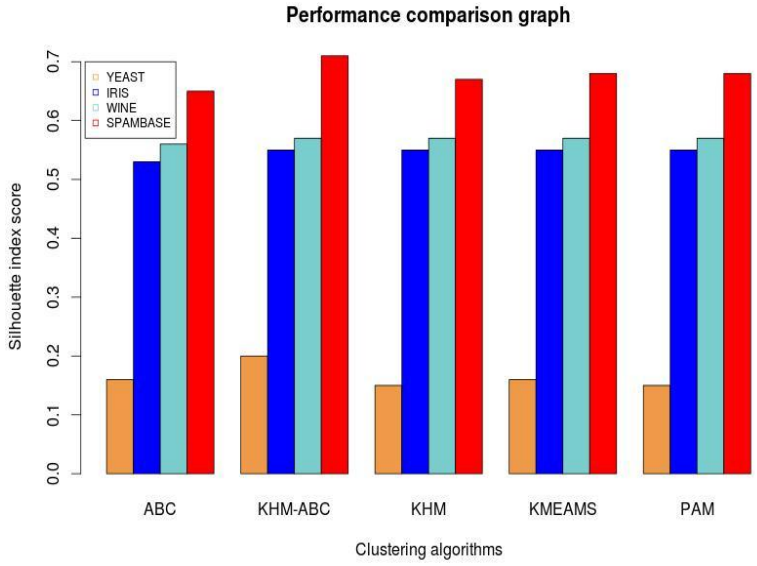

Figure 4. Performance comparison graph

REFERENCES

[1] K. A. A. Nazeer and M. P. Sebastian, "Improving the Accuracy and Efficiency of K–means clustering algorithm", Proceedings of the World Congress on Engineering 2009, vol. Vol I, 2009.

[2] K.Thangavel and N. Visalakshi, "Ensemble based Distributed K-Harmonic Means Clustering", International Journal of Recent Trends in Engineering, vol. Vol 2, NO.1, pp 125–129, 2009.

[3] G. Armano and M. R. Farmani, "Clustering Analysis with Combination of Artificial Bee Colony Algorithm and k-means technique", International Journal of Computer Theory and Engineering, Vol 6, Part 2, pp 141-145, 2014

[4] S. Reyya, M. Pushpa, and et al, "Increasing Comparison Performance using K-Harmonic Mean", International Journal of Management, Information Technology and Engineering, vol. Vol 2, Issue 3, pp 11–18, 2014.

[5] N. Alldrin, A. Smith, and D. Turnbull, "Clustering with EM and K-Means", Department of Computer Science, University of California, San Diego

[6] B. Zhang, M. Hsu, and U. Dayal, "K-Harmonic Means - A Data Clustering Algorithm", Software Technology Laboratory,HP Laboratories Palo Alto, HPL– 1999-124, 1999.

[7] B. Zhang, "Generalized K-Harmonic Means – Dynamic Weighting of Data in Unsupervised Learning", Hewlett-Packard Laboratories, 1999.

[8] F. Nie, T. Tu, and et al, "K-Harmonic Means Data Clustering with PSO Algorithm", in Advances in Electrical Engineering and Automation, AISC, Springer Verlag, 2012, pp. 67–73.

[9] R. Jain, "A Hybrid Clustering Algorithm for Data Mining", School of Computer Science IT, Indore, India,

[10] C. Zhang, D. Ouyang, and J. Ning, "An Artificial Bee Colony approach for Clustering", Expert Systems with Applications, Elsevier, pp. 4761–4767, 2010.

[11] D. Karaboga and C. Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", Ap-plied Soft Computing, Elsevier, pp. 652–657, 2011.

[12] B. A. Dervis Karaboga, "A comparative study of Artificial Bee Colony algorithm", Applied Mathematics and Computation, Elsevier, pp. 108–132, 2009.

[13] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the Number of Clusters in a Dataset via Gap Statistic", J.R.Statist.Soc.B, Sanford University USA, vol. Vol 63, Part 2, pp. 411–423, 2001.

[14] J. Rahnenfuhrer and F. Markowetz, "Exploratory Data Analysis — Clustering Gene Expression Data", Practical DNA Microarray Analysis, Saarbrucken, 2005.

AUTHORS

Sithara E. P obtained herB.Tech (Computer Science and Engineering) from College of Engineering Vadakara, and M.Tech (Computer Science and Engineering) from NIT Calicut. She is currently working as Assistant Professor in Computer Science and Engineering Department, College of Engineering Vadakara. Her areas of interest include Bioinformatics, Data Mining, Data structures and algorithm analysis.

K. A Abdul Nazeer obtained his B.Tech (Computer Science and Engineering) from TKM College of Engineering, Kollam, University of Kerala, M.Tech (Computer Science and Engineering) from IIT Madras and Ph. D from NIT Calicut (Thesis Title: Improved Clustering Algorithms for Bioinformatics Data Analysis). He is currently Associate Professor and Head of Computer Science and Engineering Department, NIT Calicut. His areas of Interest includes