A Two-Stage Genetic *K*-harmonic Means Method for Data Clustering

Anuradha D. Thakare and Chandrashkehar A. Dhote

Abstract. Clustering techniques are aimed to partition the entire input space into disconnected sets where the members of each set are highly connected. K-harmonic means (KHM) is a well-known data clustering technique, but it runs into local optima. A two stage genetic clustering method using KHM (TSGKHM) is proposed in this research, which can automatically cluster the input data points into an appropriate number of clusters. With the best features of both the algorithm, and TSGKHM the first stage overcomes the local optima and results in optimal cluster centers, and in the second stage, results into/in optimal clusters.

The proposed method is executed on globally accepted, four real time data sets. The intermediate results are produced. The performance analysis shows that TSGKHM performs significantly better.

Keywords: Clustering, Genetic Algorithm (GA), K-harmonic means, and the TSGKHM algorithm.

1 Introduction

Cluster analysis is an important and primitive research field, which measures the identical relationship among the objects in the absence of prior knowledge.

Anuradha D. Thakare
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering,
Pune, India
e-mail: adthakare@yahoo.com

Chandrashkehar A. Dhote
Department of Computer sc. and Engineering,
Prof. Ram Meghe Institute of Technology and Research, Badnera,
Amravati, India
e-mail: vikasdhote@rediffmail.com

El-Sayed M. El-Alfy et al. (eds.), *Advances in Intelligent Informatics*,
Advances in Intelligent Systems and Computing 320, DOI: 10.1007/978-3-319-11218-3_36

K-means is a centroid based clustering method [1], which classifies given input objects in to k clusters and for many applications; it was used effectively for producing the clusters [2], and becomes popular for its feasibility. It has the ability to deal with a large amount of data, but its computational complexity becomes high. Also, due to the random choice of initial centroids, it easily runs into local optima. Several modifications are done by the researchers to improve the K-means clustering algorithm. The improved version of K-means was proposed called K-harmonic means [3, 4]. The KHM is not sensitive to the selection of initial seeds as centroids like the k-means algorithm because the clustering objective of KHM is to minimize the harmonic average from all the data points to all the cluster centers [5, 6]. The improvements in the KHM are suggested and various hybrid models are proposed by the researchers. Both K-means and KHM easily run in to local optima.

GA is a stochastic general search method [7], capable of effectively exploring the large search spaces. It is widely used because of its abilities of self-adaptation and self-organization. Mostly, it is used to solve some complicated optimization problems. The global optimization technique called the genetic algorithm is used to produce the optimal clusters. The GA has the capability to automatically divide the input space into many regions and then through a number of generations, produces accurate clusters. The quality of the solutions that GA has showed in the different types of fields and problems it makes perfect sense to try to use it in clustering problems [8]. GA minimizes the square error of the cluster of dispersion. The complex problems such as unsupervised clustering or non-parametric clustering are often dealt with by employing an evolutionary approach.

A method called the genetic K-means algorithm (GKM) was proposed [9], which defines a distance-based mutation. This is a basic mutation operator specific to clustering.

A new hybrid data clustering algorithm based on KHM and improved GSA, called IGSAKHM, was proposed [10], which not only helps the KHM clustering escape from local optima, but also overcomes the slow convergence speed of the IGSA.

A hybrid algorithm for clustering called PSOKHM was proposed [11], which uses Particle Swarm Optimization to help KHM escape from local optima at a certain level, and results in better clustering. IGSAKHM is superior to the KHM algorithm and the PSOKHM algorithm in most cases.

In this paper, we have proposed a two stage genetic *k*-harmonic means method for data clustering, which overcomes the problem of local optima by integrating the clustering objective function of KHM as a fitness function in GA. In the first stage, GA results in globally optimal cluster centers and in the second stage, GA uses the KHM function for cluster formation. The proposed *TSGKHM* method is found to work very satisfactorily.

1.1 K-Harmonic Means Clustering

KHM uses the Harmonic Averages of the distances from each data point to the centers as it is constituent to its performance function [10]. The K-Harmonic Means algorithm overcomes the drawbacks of K-Mean in terms of initialization of centers. KHM uses different objective functions [10] one of, which is as follows:

$$KHM = \sum_{k=0}^{n} \left(\frac{\text{K cluster centers}}{\left(\sum \textit{average} \text{ of distance from all points to cluster center}\right)^{p}} \right) \quad (1)$$

Where, P is an input parameter.

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

2 Related Work

A self-organized genetic algorithm for document clustering was proposed [12] based on semantic similarity measure. The Genetic Algorithm in conjunction with the hybrid strategy, gets the best clustering performance. The self-organized genetic algorithm, considering the influence between the diversity of the population and the selective pressure, efficiently, evolves the clustering of the documents in comparison with the standard k-means algorithm in the same similarity strategy.

Another hybrid approach is a combination of K-harmonic means and the Particle Swarm Optimization (PSOKHM) algorithm. The PSO method is a population based global optimization technique in, which the solution space of the problem is expressed as a search space. Each position in the search space is an interrelated solution of the problem [5]. The KHM algorithm tends to converge faster than the PSO, but the drawback is that it gets stuck in local optima. The hybrid clustering algorithm called PSOKHM [11], maintains the qualities of KHM and PSO. The 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 a collection of particles. To improve fitness value, the Particle Swarm Optimization algorithm is applied for eight generations.

Various clustering approaches using a center-based clustering algorithm, KHM have been proposed by the researchers. The main objective of the clustering is to produce the accurate and effective clusters with less computation time. The performance of the KHM [13], is measured based on a function, which calculates Harmonic averages of the distances from each data point to the centers. The objective function in KHM, computes the membership function and weight associated to/with each of the data points. The centers location is recalculated from all the data points having the highest membership value, to assign a data point to the cluster. A two-stage genetic clustering algorithm can automatically, determine the

proper number of clusters and the proper partition from a given data set [14]. The two-stage selection and mutation operations are implemented to exploit the search capability of the algorithm.

In the Gravitational Search Algorithm, the objects are considered to be with masses, which attract each other by the gravity force. All the objects move towards the ones with heavier masses due to force. The information is being transformed by the gravitational force among the objects and the objects, which have a higher mass become heavier. The Gravitational Search Approach using the K-Harmonic Means method called GSAKHM [10] was proposed, which is an effective way of clustering the documents and can be achieved by using a combination of the Gravitational Search Method and the K-Harmonic Means algorithm.

The hybrid clustering method using two popular clustering techniques, ant based clustering and the K-Harmonic Means algorithm (ACAKHM) was proposed [15]. ACA works on the principle of an ant's behaviour of collecting or organizing the feedback of the behaviour of the ants. Even though, ACA provides appropriate partitions without defining the initial cluster centers, it takes a long time to get a better result. To overcome this drawback, ACAKHM makes a better use of the advantage of both ACA and KHM. ACA is effectively used to get the initial canters and then KHM is used to avoid local optima.

Candidate group search is based on some selection rules to isolate the candidate groups for each center (CGSKHM). Screening through all the data sets. If it fits in to the candidate group, the center has to be interchanged and a new solution is achieved by using KHM. The Candidate Group Search [16], offers a scheme combining of some haphazardness and deterministic selection rules coming from the data set, CGS outperforms KHM and it requires less computation time. Variable neighbourhood search (VNS) is a meta-heuristic technique intended to resolve combinatorial and global optimization problems. Variable neighbourhood search for harmonic means clustering (VNSKHM) was proposed [17] in, which, the Variable neighbourhood search is experimented to solve the problem of KHM clustering, which is easily trapped in local optima. The elementary idea is to continue to a systematic change of neighbourhood within a local search algorithm.

3 Proposed Method

The KHM algorithm is found to be a good clustering algorithm for many applications. The best feature is that , it requires less function evaluations and hence, converge is faster. But it gets stuck into local optima most of the times and therefore, the technique with global search abilities is required to improve the performance.

The proposed method integrates the KHM and GA to produce optimal clusters, called Two-Stage Genetic KHM. The best features of both the techniques are used for cluster formation. The input space contains the clusters on, which the proposed two-stage GA is applied for cluster optimization in two stages with two objective functions, one for each stage. In the first stage, the clustering objective minimum

intracluster distance is used as the fitness function in a genetic process, which results in subclusters formation. In the second stage, the KHM function is used as a fitness function for cluster optimization. The steps carried out are as given in the algorithm.

Algorithm: ***TSGKHM***
Steps:

1. Initial parameter setting with number of generations and population size
2. Generate the population
3. Fitness evaluation with *fitnessfun_1*
4. Apply GA operators & generate a new population
5. Repeat step 3 & 4 till convergence
6. Store Sub clusters with their fitness values
7. Fitness evaluation with *fitnessfun_2*
8. Apply GA operators & generate a new population
9. Repeat step 7 & 8 till convergence
10. Optimal clusters

3.1 Fitness Function_1: Intracluster Distance [18]

In the first stage GA, the fitness function is intracluster distance *E*. The data points with minimum *E* value are the most fitted members for the cluster formation. On this basic theme, the subclusters are formed. Suppose a data point x_i is assigned to the j[th] cluster based on minimum euclidean distance. The intra-cluster distance [19] is calculated by following formula:

$$E = \sum_{x_i \subseteq c_j} \frac{\left\| x_i - x_j \right\|}{N_j} \qquad \text{for } i = 1, 2, 3, \ldots N \tag{2}$$

Where, x_j is the center of cluster c_j, and *N* is the size of population i.e. number of data points.

3.2 Fitness Function_2: (KHM) K-Harmonic Means [10]

The K-Harmonic Mean of the data point, is calculated by equation 1.

$$KHM(X, C) = \sum_{i=1}^{m} - \frac{n}{\sum_{j=1}^{n} \frac{1}{\| x_i - c_j \|^p}} \tag{3}$$

Where, *N* is number of clusters, x_i is number of objects, and c_j is cluster center. In order to evaluate the membership of a data point, the function [10] in equation 2 is used.

$$m\left(\frac{C_j}{X_i}\right) = -i\left(\frac{\| X_i - C_j \|^{-p-2}}{\sum_{j=0}^{k} \| X_i - C_j \|^{-p-2}}\right) \tag{4}$$

The center is recalculated in the KHM algorithm [10] by using the formula given below.

$$C_j = \frac{\sum_{i=1}^{n} m\left(\frac{C_j}{X_i}\right) w(X_i) X_i}{\sum_{i=1}^{n} m\left(\frac{C_j}{X_I}\right) w(X_i)} \tag{5}$$

4 Results and Discussion

4.1 Dataset Used

The experimentation results, initially, four globally accepted real life data sets are taken. These data sets are described in terms of the number of points present, dimensions, and the number of clusters. These four real life data sets are taken from a UCI machine learning repository [19], which represents examples of data with low, medium, and high dimension. The data set with this description are necessary for calculating the objective function values of clusters. Initially we have tested the results of our proposed work on selected four datasets in future we may test it on more datasets.

(a) *Fisher's Iris* dataset *(n=150, d=4, k=3)*: It contains three different species of iris flower and 50 samples were collected from the four features that are: sepal length, sepal width, petal length, and petal width.
(b) *Wine* dataset *(n=178, d=13, k=3)*: It consists of 178 objects characterized by 13 features. These features are obtained by a chemical analysis of wines that are produced in the same region in Italy.
(c) *Cancer* dataset *(n=683, d=9, k=2)*: It contains 683 objects characterized by nine features. All the data is categorized in to two classes, malignant tumors (444 objects) and benign tumors (293 objects).
(d) Glass dataset (*n = 214, d = 9, k = 6*): It contains 214 objects characterized by nine features. The objects are categorized into six classes.

4.2 Performance Measures

In order to evaluate the performance of a proposed method we have used two measures i.e. KHM and *F*-measure. The runtime is also recorded for all the methods. A measure, which is used to evaluate the performance of a classification model [20, 21] where all the class labels are known is used. *F*-measure is the measure of the test of accuracy with an optimum score as 1 giving good clustering and combines both precision and recall to compute the scores. Precision is the

fraction of retrieved instances that are relevant. Recall is the fraction of retrieved instances that are relevant. KHM is the summation of all data points of the harmonic average of the distance from a data point to all the centers, as defined in equation (3). The smaller value of sum indicates higher quality of clustering.

4.3 Discussion on Results

In order to form good quality clusters, the proposed work is divided into two stages. In the first stage, the sub clusters are formed using GA. The objective function, intracluster distance is used as a fitness function for GA. Smaller the values of intracluster distance, higher the quality of clusters. The clusters formed are considered as subclusters so that final clusters can be formed in the second stage, using the KHM as a fitness function for GA. The best results are tabulated for each data set. Table 4.1, shows the results of four datasets in terms of sub-clusters formed in the first stage. These subclusters are nothing, but the actual clusters formed by *fitness function_1*. The input parameter P, *F*-measure, and runtime values are taken from the second stage calculations. *F*-measure values represent the clustering accuracy. These values change with the change in input. The runtime increases with respect to the size of the dataset and input parameter values.

Table 4.1 Results of proposed *TSGKHM*

Performance Criteria	Datasets			
	Iris	**Wine**	**Cancer**	**Glass**
Size of dataset	150	178	683	214
Sub-clusters	9	8	12	5
Inputs (P)	4	2	4	2
***F*-measure**	0.89	0.86	0.65	0.89
Runtime	78	1950	3806	3463

The second stage results for each dataset are recorded in the following tables. Table 4.2 shows the intermediate results for the various datasets. The subclusters resulted from the first stage and the number of data points in each subcluster is tabulated. The KHM values for some subclusters of iris dataset are exceeded even ~600 because these are the intermediate results and the subclusters will be more refined further at the time of final cluster formation. When these subclusters are given as input, in the second stage, we got the results in terms of KHM and *F*-measure. The KHM values are varying for each subcluster because of the number of data points present. Also, we got minimum values of KHM for some subclusters, as the subclusters are already optimized using GA. The reason for using KHM is to optimize the number of clusters by subcluster merging. In order to merge the subclusters, the KHM values are compared. The KHM value reflects the quality of the subcluster based on, which the decision will be taken whether to

merge the subcluster with the other or to keep it separate. The exact logic for subcluster merging may be enhanced to improve the results.

Table 4.2 Results of proposed *TSGKHM* in terms of subclusters on the datasets

Subcluster no.	# of Data points	Runtime	KHM	F-measure
Iris Dataset for P=4				
1	01	125	0.0141	1.05
2	01	78	0.0453	0.30
3	04	63	0.1205	0.44
4	45	47	117.87	0.74
5	35	78	117.88	0.89
6	11	78	117.90	1.04
7	03	62	117.92	1.02
8	29	62	147.86	1.63
9	21	62	578.27	1.08
Wine Dataset for P=2				
1	06	2059	0.0108	1.05
2	64	2106	0.0109	0.30
3	02	1872	0.0434	0.44
4	21	1856	0.0456	0.74
5	38	1950	0.1632	0.89
6	06	2090	0.4085	0.49
7	03	1934	0.4101	0.52
8	38	1669	0.0098	0.59
Glass Dataset for P=2				
1	05	3385	0.00010	0.30
2	14	3463	0.00110	0.89
3	18	2886	0.00100	1.02
4	29	3447	0.00120	1.05
5	17	2885	0.00010	1.08
6	19	3.403	0.00214	0.28
7	22	3385	0.00031	0.27
8	24	3432	0.00045	0.37
9	18	3338	0.00160	0.36
10	20	3385	0.00010	0.32
11	08	3354	0.00821	0.35
12	20	3417	0.00032	0.45
Cancer Dataset for P=4				
1	10	3088	0.5210	0.25
2	198	3027	0.3425	0.56
3	12	2793	0.2489	0.50
4	434	2823	0.2560	0.46
5	29	3806	0.0241	0.65

Table 4.3 Comparison of intermediate results of *TSGKHM* with the existing methods PSOKHM and GSAKHM[10]. Here, KHM is the KHM function values, *F*-m is the *F*-measure, and Rt is the runtime parameter.

Dataset	Criteria	PSOKHM			GSAKHM			TSGKHM		
		KHM	*F*-m	Rt	KHM	*F*-m	Rt	KHM	*F*-m	Rt
Iris	P=4	106.06	0.751	1.967	704.48	0.8471	2.799	0.0141	1.05	125
Wine	P=2	59844	0.829	9.525	7.046	0.9440	4.461	0.0108	1.05	2059
Glass	P=2	1196.7	0.424	17.669	7.0236	0.9897	3.994	0.0001	1.08	2885
Cancer	P=4	NA	NA	NA	701.03	0.6824	2.352	0.0241	0.65	3806

The intermediate results of TSGKHM are tabulated. These will be used for analysis purposes so as to find out the exact criteria for subcluster merging. The results of the proposed work are compared with the existing hybrid KHM methods i.e. PSOKHM and GSAKHM [10] as tabulated in Table 4.3 Here, The best values for each dataset are taken for the comparison. These comparisons are made to analyze the deviations in the values of performance measures. Further, the results would be more refined after final cluster formation.

5 Conclusions

The two-stage genetic algorithm using K-harmonic means for data clustering is proposed, which can automatically cluster the entire data to produce the optimal clusters. We presented the results of the first stage in terms of subclusters using GA and results of the second stage in terms of KHM values and *F*-measure. The GA utilizes two fitness functions, intraclsuter distance to form subclusters and KHM to get the optimal number of clusters. We got the desired results for the first stage and in the second stage; the experimentation is done for calculating KHM values. The minimum values for KHM reflects good clustering hence, this observation will be used further for subclusters merging to get the final clusters.

References

[1] Macqueen, J.B.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)

[2] Jiawei Han, M.K.: Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier (2006)

[3] Zhang, B., Hsu, M., Dayal, U.: K-harmonic means – a data clustering algorithm. Technical Report HPL-1999-124. Hewlett-Packard Laboratories (1999)
[4] Zhang, B., Hsu, M., Dayal, U.: K-harmonic means. In: International Workshop on Temporal, Spatial and spatio-Temporal Data Mining, TSDM 2000, Lyon, France (September 12, 2000)
[5] Cui, X., Potok, T.E.: Document clustering using particle swarm optimization. In: IEEE Swarm Intelligence Symposium, Pasadena, California (2005)
[6] Kao, Y.T., Zahara, E., Kao, I.W.: A hybridized approach to data clustering. Expert Systems with Applications 34(3), 1754–1762 (2008)
[7] Karegowda, A.G., Manjunath, A.S., Jayaram, M.A.: Application of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis of Pima Indians Diabetes. International Journal of Computer Applications (0975 – 8887) 43(1) (April 2012)
[8] Painho, M., Bação, F.: Using Genetic Algorithms in Clustering Problems. In: GeoComputation. Higher Institute of Statistics and Information Management, New University of Lisbon, Travessa, EstêvãoPinto (Campolide) P-1070-124 Lisboa (2000)
[9] Krishna, K., Murty, M.N.: Genetic k-means algorithm. IEEE Transactions on Systems, Man and Cybernetics B Cybernetics 29, 433–439 (1999)
[10] Yin, M., Hu, Y., Yang, F., Li, X., Gu, W.: A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering. Expert Systems with Applications 38, 9319–9324 (2011)
[11] Yang, F., Sun, T., Zhang, C.: An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization (2009)
[12] Song, W., Park, S.C.: An Improved GA for Document Clustering with semantic Similarity measure. In: Fourth International IEEE Conference on Natural Computation, pp. 536–540 (2008)
[13] Zhang, B., Hsu, M.: K-Harmonic Means - A Data Clustering Algorithm Umeshwar Dayal Software Technology Laboratory HP Laboratories, Palo Alto HPL-1999-124 (October 1999)
[14] He, H., Tan, Y.: A two-stage genetic algorithm for automatic clustering. Neurocomputing 81, 49–59 (2012)
[15] Jiang, H., Yi, S., Li, J., Yang, F., Hu, X.: Ant clustering algorithm with K-harmonic means clustering (2010)
[16] Hung, C.H., Chiou, H.-M., Yang, W.-N.: Candidate groups search for K-harmonic means data clustering (2013)
[17] Alguwaizani, A., Hansen, P., Mladenovic, N., Ngai, E.: Variable neig
[18] Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. Willey series in probability and statistics
[19] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
[20] Sahaa, S., Bandyopadhyay, S.: A generalized automatic clustering algorithm in a multiobjective framework. Applied Soft Computing 13, 89–108 (2013)
[21] Handl, J., Knowles, J.: An evolutionary approach to multiobjective clustering. IEEE Transaction on Evolutionary Computation 11(1), 56–76 (2007)