On K-means Data Clustering Algorithm with Genetic Algorithm

Shruti Kapil
Dept. of Computer Science Engineering,
Giani Zail Singh Campus,
Bhatinda, India
kapilshruti31@yahoo.com

Meenu Chawla
Dept. of Computer Science Engineering,
Giani Zail Singh Campus,
Bhatinda, India
meenuchawla011@gmail.com

Mohd Dilshad Ansari
Dept. of Computer Science Engineering,
Jaypee University of I.T.,
Waknaghat, India
m.dilshadcse@gmail.com

***Abstract*— Clustering has been used in various disciplines like software engineering, statistics, data mining, image analysis, machine learning, web cluster engines, and text mining in order to deduce the groups in large volume of data. The notion behind clustering is to ascribe the objects to clusters in such a way that objects in one cluster are more homogeneous to other clusters. There are variegated clustering algorithms available viz k-means clustering, cobweb clustering, db-scan clustering, fartherstfirst clustering, and x-means clustering algorithm but K-means on the whole comprehensively used algorithm for unsupervised clustering dilemma. In this paper k-means clustering is being optimised using genetic algorithm so that the problems of k-means can be overridden. The outcomes of k-means clustering and genetic k-means clustering are evaluated and compared; obtained result shows K-means with GA algorithm suggest new improvements in this research domain.**

Keywords: Data mining, K-means clustering, Genetic algorithm, Euclidean distance, Crossover, Mutation, Genetic k-means algorithm.

I. INTRODUCTION

Clustering is the unsupervised method used to segregate the data into clique so that objects belonging to one group are similar and different from the objects in other groups [12, 14]. A high-quality clustering practice is that which provide high homogeneity within cluster and heterogeneity among the clusters [10, 14]. K-means clustering is the method of partitioning in which objects are grouped into user defined 'K' number of clusters [8]. This method aims at optimizing the cost functions to minimize the intera-cluster disparity and to maximize the inter-cluster disparity.

Genetic algorithm which proposed early in 1989 [5, 9] is search heuristic usually applied in the optimization problems [13] guided by the principles of evolution and natural genetics [1-8]. Genetic algorithm belongs to the larger class of evolutionary algorithms which engender solutions to optimization problems using techniques inspired by natural evolution such as inheritance, mutation, selection and crossover [10] .In GA, the populace of candidate solutions to an optimization problem is evolved towards better solutions [6, 15]. Each candidate solution has properties which can be Mutated and altered [15]. The basic elemental artistry of the GAs are designed to mimic processes in natural systems necessary for evolution, the principle stated by Charles Darwin "Survival of Fittest"[1,6]. GAs imitates the survival of the fittest individual over successive generations for solving any problem [1]. Each generation consist of string of characters which are similar to the chromosome of the DNA. Each individual is the possible solution of the problem basis of fitness of each individual the individual with maximum fitness forms the solution of the problem. The GAs is analogous to the biological genetic structure [9].

- Individuals in a population race for resources and mates [1].
- The individuals who won the race will reproduce to form more offspring as compared to those who perform poorly.
- Each consecutive generation will become more adaptive to their atmosphere.

The rest of the paper is organized as follows. The next section provides a detailed over view of K-means based data clustering. Section 3 contains the details of GA. The section 4 explains GA based data clustering. Results and conclusion are described in the section 5 and 6 respectively.

II. K-MEANS CLUSTERING

The K-means clustering is unsupervised algorithm used to clique different object into clusters. A cluster is collection of data objects that are homogeneous within one cluster and heterogeneous to object in other cluster [13]. A cluster of data object can be treated collectively as one group and so may be considered as a form of data compression [7]. The K-means Clustering algorithm has the following steps:

1) Label the number of clusters.
2) Establish the centroid coordinate.
3) Determine the distance of each object to the centroid.
4) Group the objects based on minimum distance.

978-1-5090-3669-1/16/$31.00 ©2016 IEEE

K-means Clustering is partitioning method of grouping 'n' observations into 'k' clusters basis of minimum distance between centre of cluster and the observation point [13].

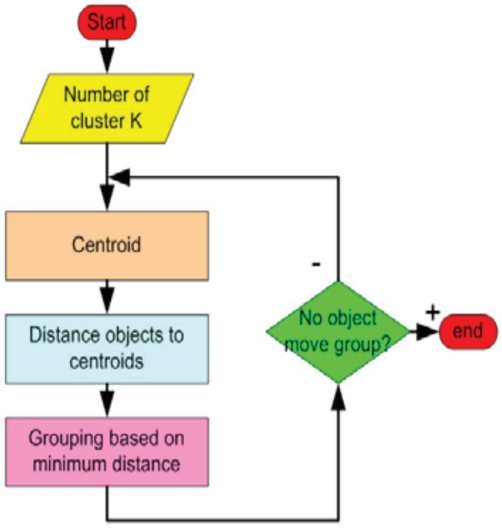

Fig.1. Steps of K-means Clustering.

It is conducted iteratively so that observation point is at least distance from cluster centre [2] . K-means clustering uses various distance function to measure the similarity among the objects. The distance functions used by algorithm are Euclidean distance metric function and Manhattan distance metric function.

Although k-means is simple and can be used for a wide variety of data types [14], but it suffers from some drawback as k-means algorithm is computationally dear and requires more time which is relative to number of data items, number of clusters and number of iterations [6, 14], need to specify the number of clusters beforehand.

III. GENETIC ALGORITHM

Genetic algorithm works on the notion of coding parameter sets rather than parameters themselves [13]. The encoded parameter sets are known as chromosomes. GAs computes the optimization problems using population of fixed size known as population size [13]. A solution consists of string of symbols quintessential binary symbols. The more fit members of this population are more likely to mate and produce the next generation. As the generation pass, the members of the population get closer and closer to the solution.

Outline of Genetic Algorithm:

Generate 'P' be randomly set of solutions for an embryonic generation.

Step 1. t=0

Step 2. While t<T or termination criteria not meet do

Step 3. Compute the fitness factor of P (t).

Step 4. Select P_b the fitted solution for next generation.

Step 5. t=t+1

Step 6. Perform crossover to generate new solution.

Step 7. Perform mutation to the solutions.

Step 8. End while.

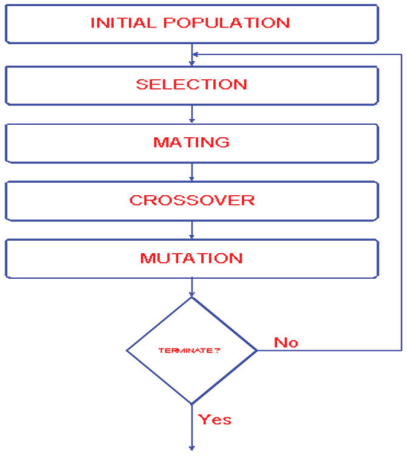

Fig.2. Steps of Genetic Algorithm [13]

The genetic algorithm is stimulated by the organic progression and the operators like Natural Selection, Crossover, and mutation mimic the process of biological evolution.

IV. GENETIC K-MEANS CLUSTERING

The basic steps of GAs are also followed in GA-clustering algorithm [6]. The genetic operators that are used in Genetic K-means Algorithm are selection, the distance based mutation and k-means operator [13]. The following section explains the genetic k-means clustering algorithm by specifying coding, initialization schemes and the genetic operators.

1) *Initialization:* The data objects acts as candidate for the centre of cluster. The length of chromosome will be similar to the range of the data set. The n^{th} gene of the chromosome corresponds to the n^{th} data point in the data set. If any data point is the centre of cluster then the allele value of the chromosome will be '1' otherwise '0' [5].

2) *Fitness computation:* In this phase all the instance has been evaluated for computation of fitness value based of fitness function provided by the system. On the basis of this fitness function fitness for the entire instance has been computed. The fitness of the instance is calculated in terms of inter-dependency among them.

3) *Selection*: This Process selects the chromosome from the mating pool basis of the Charles Darwin Principle "Survival of Fittest". In the proportional selection strategy adopted in this article, a chromosome is assigned a number of copies, which is proportional to its fitness in the population that go into the mating pool for further genetic operations.

4) *Crossover:* This Process swap over properties of the two individuals to produce the offspring. Selected mates should have properties to fit into next generation so that crossover may produce good individuals [7].

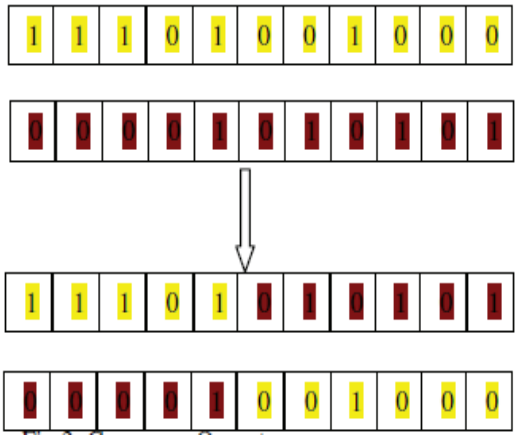

Fig.3. Crossover Operator

5) Mutation: In this process the allele value of the chromosome get mutated basis of fixed mutation probability. For binary chromosome the mutation is done by simply flipping the allele value i.e. '0' to '1' or '1' to '0'.

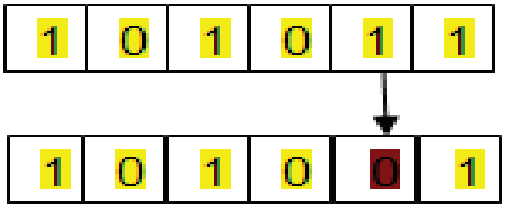

Fig.4. Mutation Operator

6) *K-means operator:* Thus the fittest instance is being selected as the centroid to further perform k-means clustering.

V. EXPERIMENTAL RESULTS

K-means and Genetic K-means clustering algorithm is being carried out to evaluate the performance of k-means and genetic k-means algorithm. All the experiments are carried on dummy data whose description is as follows:
The online user data set containing about the 200 instances and their activities. This data set contains various attributes which explain about their online behaviour about any specific social networking site.

TABLE 1: ATTRIBUTES USED TO CLASSIFY USERS

GROUP	ATTRIBUTES
Status attribute	Number of friends
	Number of followers
Login attributes	Number of login times
	Number of login days
	Number of active times
	Number of active days
	Number of total online time
Basic operation attributes	Number of blogs written
	Number of pictures uploaded
	Number of videos shared
	Number of time password changed
	Number of status updated
	Number of friend request sent/received
	Number of messages in inbox
	Number of notifications
	Number of pages like
	Number of times profile updated
	Total number of basic operations
Application operation attribute	Number of times of playing game 'A'
	Number of times of playing game 'B'
Activity Users	Number of login times
	Number of login days
	Number of messages

To implement this clustering algorithm 'WEKA' is used. WEKA is the product of University of Waikato and supports various data mining functions like data pre-processing, clustering, classification, visualisation, regression and feature selection. The software is written in the Java language and contains a GUI for interaction with data files and producing visual results [15].

The Figure 5 represents the K-means clustering performed with k=2 upon supplied data set using Euclidean distance function.

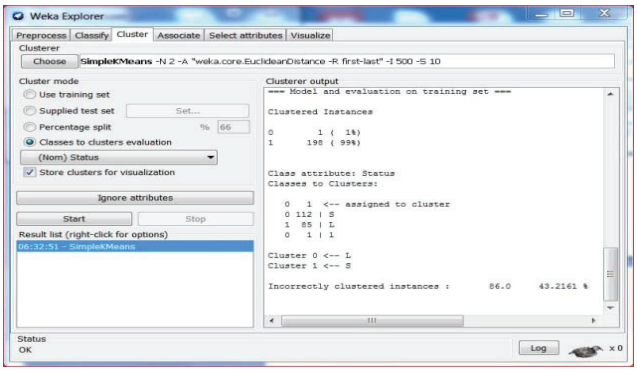

Fig.5.Simple K-means Clustering with k=2 using Euclidean distance.

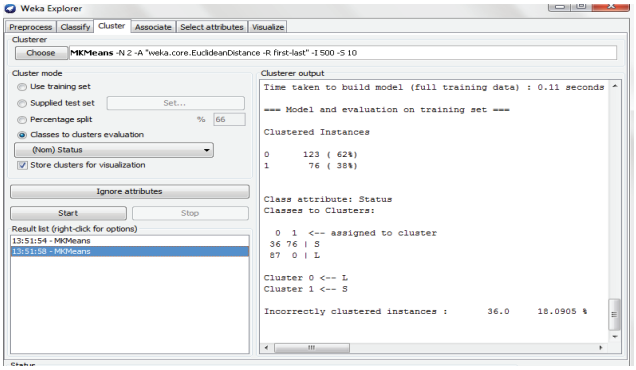

Fig.6. K-means clustering with GA.

The Figure 6 represents the K-means clustering performed with k-means with Genetic Algorithm using k=2 and Euclidean distance.

The following figures shows the comparison between the performance of k-means clustering and k-means clustering with genetic algorithm in terms of accuracy, incorrectly clustered clusters and sum squared errors.

$$\text{Accuracy } (\%) = \frac{\textit{Total Correctly Clusterd Instances}}{\textit{Total Instances}} * 100$$

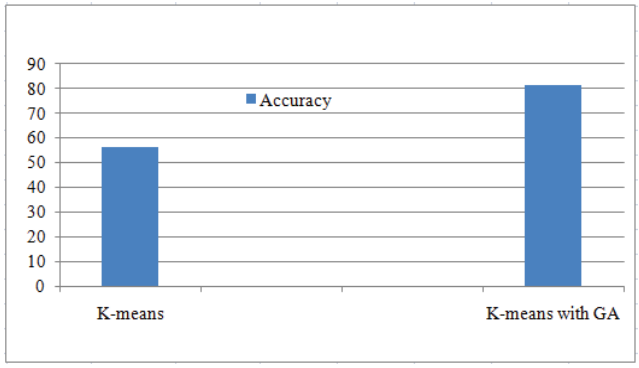

Fig.7.Accuracy Graph of Different Clustering Approaches.

The figure 7 represents graphical representation of accuracy achieved during clustering by different algorithm. Accuracy has been computed by using ratio of total correctly classified instance to total instances available in dataset.

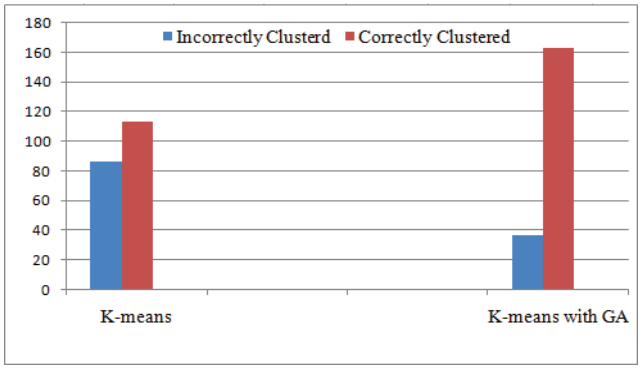

Fig.8. Incorrectly and Correctly Clustered Instances.

Figure 8 represents instances that have been correctly clustered and that are incorrectly clustered by different Clustering algorithms. Higher the correctly cluster instances represents much reliable clustered technique is used for clustering.

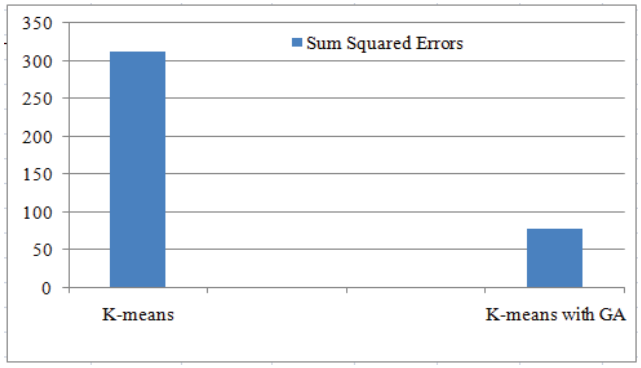

Fig. 9.Sum Squared Errors.

Figure 9 represents errors in Clustering of instances that have been correctly and incorrectly clustered by different clustering algorithms. Sum Squared Error is the sum of squared difference between observation and its group's mean. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the Sum Squared Error would be equal to 0.

The Sum Squared Error is being Calculated as:-

$$SSE = \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

Where n is the number of observations x_i is the value of the $\boldsymbol{i^{th}}$ observation.

VI. CONCLUSION

K-means clustering is unsupervised algorithm used to form different clusters of a data set so that similar data are grouped together. The K-means clustering is although widely used with various data types but it has some drawbacks which make it infeasible. So to overcome these problems genetic algorithm is used along with k-means clustering to produce the fittest result. The results show that the genetic K-means clustering outperforms the k-means clustering in terms of sum squared errors and correctly clustered instances. In future, the performance of different clustering algorithm along with different machine learning algorithm for different distance metric can be figured out using other high dimensional data sets.

REFERENCES

[1] D.E. Goldberg "Genetic Algorithms in Search Optimization and Machine Learning", Addison-wesley, New York-1989.

[2] L. Davis (Ed.), Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, 1991.

[3] Z. Michalewicz "Genetic Algorithms Data Structure" Evolution Programs, Springer, New York, 1992.

[4] Ribeiro Filho, José L., Philip C. Treleaven, and Cesare Alippi. "Genetic-algorithm programming environments." Computer 27, Vol. 6, pp. 28-43. 1994.

[5] Pal, Sankar K., Dinabandhu Bhandari, and Malay K. Kundu. "Genetic algorithms for optimal image enhancement." Pattern Recognition Letters, Vol.15 (3), pp. 261-271, 1994.

[6] Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. "Genetic algorithm-based clustering technique." Pattern recognition, Vol.33 (9), pp.1455-1465, 2000.

[7] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, San Francisco, USA,2001.

[8] Chiou, Yu-Chiun, and Lawrence W. Lan. "Genetic clustering algorithms." European journal of operational research, Vol.135 (2), pp. 413-427, 2001.

[9] Zheyun Feng "Data Clustering using Genetic Algorithm" Evolutionary Computation: Project Report, CSE484, 2012.

[10] Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." IEEE Transactions on neural networks, Vol.16(3), pp. 645-678, 2005

[11] Krantz, Amanda, Randi Korn, and Margaret Menninger. "Rethinking Museum Visitors: Using K-means Cluster Analysis to Explore a Museum's Audience.", Curator: The Museum Journal. Vol. 52(4), pp.363-374, 2009.

[12] Wei, Chih-Ping, and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." *Expert* systems with applications, Vol.23(2), pp.103-112, 2002

[13] Dash, B., Mishra, D., Rath, A., & Acharya, M., "A hybridized K-means clustering approach for high dimensional dataset", International Journal of Engineering, Science and Technology, Vol.*2* (2), pp.59-66, 2010.

[14] Sharma, Sonia. "ShikhaRai: Genetic K-means algorithm-implementation and analysis." Int. J. Recent Technol. Eng. (IJRTE), Vol.1 (2), 2012.

[15] Dash, Rajashree, and Rasmita Dash. "Comparative analysis of k-means and genetic algorith based data clustering." International Journal of Advanced Computer and Mathematical Sciences, Vol.3 (2), pp.257-265, 2012.

[16] Singh, Archana, Avantika Yadav, and Ajay Rana. "K-means with Three different Distance Metrics." International Journal of Computer Applications, Vol. 67(10), 2013.

[17] Sinwar, D., and R. Kaushik. "Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering." International Journal for Research in Applied Science and Engineering Technology (IJRASET) \, Vol.2 (5), pp.270-274, 2014.

[18] Sonia Sharma, and Shikha Rai "Genetic K-means Algorithm-Implementation and Analysis", International Journal of Recent Trends in Engineering, Vol.1 (1), pp.1-4, 2014.

[19] Ansari, Mohd Dilshad, S. P. Ghrera, and Vipin Tyagi. "Pixel-based image forgery detection: a review", IETE Journal of education, Vol.55(1), pp.40-46, 2014.

[20] M.D. Ansari, G. Singh, A.Singh and A.Kumar,"An Eff cient Salt and Pepper noise Removal and Edge preserving Scheme for Image Restoration", Int.J.Computer Technology & Applications, Vol.3(5), pp.1848-1854, 2012.

[21] Verma, Gunjan, and Vineeta Verma. "Role and applications of genetic algorithm in data mining." International journal of computer applications, Vol. 48(17), pp.5-8, 2012.

[22] Ansari, M.D. and Ghrera, S.P. and Wajid, M."An Approach for Identication of Copy-Move Image Forgery based on Projection Prof ling", Pertanika Journal o f Science & Technology, In Press.

[23] Ansari, Mohd Dilshad, Ghrera, Satya Prakash and Mishra, Aruno-daya,"Texture Feature Extraction Using Intuitionistic Fuzzy Local Binary Pattern", Journal of Intelligent System, In Press.

[24] Surbhi Aggarwal, Neena Madan "Comparison Between various Approaches for Customer Relationship Management in Data Mining" International Journal of Engineering and Computer Science, Vol.5(4), pp.16257-16262, 2016.

[25] Ansari, M.D. and Ghrera, S.P."Intuitionistic fuzzy local binary pattern for features extraction", Int. J. Information and Communication Technology, In Press.

[26] Shruti Kapil, Meenu Chawla, "A Review: Comparative Analysis of Data Mining Algorithm for Identification of Numb Users", International Journal of Computer Application, Vol. 6(2), pp. 77-84, 2016.