International Journal of Computer Applications Technology and Research
Volume 6–Issue 8, 384-390, 2017, ISSN:-2319–8656

Comparative Analysis of Hybrid K-Mean Algorithms on Data Clustering

Navreet Kaur
Department of Computer Science and Engineering
Sri Guru Granth Sahib World University
Fatehgarh Sahib, India

Shruti Aggarwal
Department of Computer Science and Engineering
Sri Guru Granth Sahib World University
Fatehgarh Sahib, India

Abstract: Data clustering is a process of organizing data into certain groups such that the objects in the one cluster are highly similar but dissimilar to the data objects in other clusters. K-means algorithm is one of the popular algorithms used for clustering but k-means algorithm have limitations like it is sensitive to noise ,outliers and also it does not provides global optimum results. To overcome its limitations various hybrid k-means optimization algorithms are presented till now. In hybrid k-means algorithms the optimization techniques are combined with k-means algorithm to get global optimum results. The paper analyses various hybrid k-means algorithms i.e. Firefly, Bat with k-means algorithm, ABCGA etc. The Comparative analysis is performed using different data sets obtained from UCI machine learning repository. The performance of these hybrid k-mean algorithms is compared on the basis of output parameters like CPU time, purity etc. The result of Comparison shows that which k-means hybrid algorithm is better in obtaining cluster with less CPU time and also with high accuracy.

Keywords: Data mining, Clustering, Hybrid K-means Algorithm, ABCGA, CPU time

1. INTRODUCTION

Data mining [1] is a powerful concept for data analysis and defined as the discovery of hidden pattern from data sets. It is also defined as the Mining of knowledge from huge amount of data. Data Mining was developed to make useful discoveries from the data independently, without depending on the statistics. It is an important subfield of the computer science. The goal of the data mining process is to discover the interesting patterns from the data sets and then transforming them into a structure which is understandable for further use. Data mining is the analysis step of the knowledge discovery in databases process or KDD. Data mining is the extraction of patterns and knowledge from large amounts of data, not the extraction of data itself. It is easily applied to any form of large-scale data or information processing as well as any application of computer decision support system, including artificial intelligence, machine learning and business intelligence.

1.1 Clustering

Clustering [2] is the process of divide the population or data points into different groups such

that data points in the same groups are more similar to other data points in the same group than those in other groups. Clustering can be considered the most important unsupervised learning problem. A loose definition of clustering could be "task of organizing objects into groups whose members are similar in some way". Clustering can be said as identification of similar classes of objects. The various types of clustering methods [3] are given below:

1.1.1 Partitioning Methods

The most fundamental version of cluster analysis is partitioning, which organizes the object of dataset into groups or clusters. Some commonly used Partitioning Based clustering techniques are k-means, k-medoids.

1.1.2 Hierarchical methods

Hierarchical based clustering method group the data objects into a hierarchy or tree of clusters. Representing data objects in the form of a hierarchy is useful for data summarization and visualization. These methods are used to find spherical-shaped clusters. For example: Divisive and alggormative.

1.1.3 Density based methods

To find the clusters of arbitrary shape, this technique can model clusters as dense regions in the data space separated by sparse regions this is basic technique for density based clustering. For Example: DBSCAN (Density –based clustering based on connected regions with high Density)

1.1.4 Grid-based methods

The above clustering techniques are data-driven but the grid base clustering method takes space-driven approach by partitioning the embedding space into cells independent of distribution of input objects. For Example: STING (Statistical Information Grid).

1.2 K-means clustering

K-means clustering [4] is popular for cluster analysis in data mining. K-means clustering aims to partition m data elements into k clusters in which each data element belongs to the cluster with the minimum distance between them and which results in a partitioning of the data elements into clusters.

K-means clustering is a type of unsupervised learning when you have unlabeled data. The algorithm works iteratively to assign each data element into one of K groups based on the features that are provided. Data elements are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K clusters, they are used to label new data
2. The Labels for the training data (each data point is assigned to a single cluster)

1.2.1 Steps for k-means algorithm

Input: Number of clusters to be formed, k and a database Y= {y1, y2, y3 ...yn} containing n data elements.
Output: A set of k clusters

Method:

1. *The numbers of clusters k to be formed are chosen.*
2. ***T**he centroids are selected randomly.*
3. *The distance between each data element and cluster centroid is calculated.*
4. *The data element is assigned to the cluster centroid where distance between cluster centroid and data element is minimum than other cluster centroids.*
5. *Calculate the new cluster centroid of the data element for each cluster and update the cluster centroid.*
6. *Repeat from third step if data element was reassigned otherwise stop.*

In k-means algorithm user need to specify the number of cluster to be formed in advance and also k-means algorithm converges to local minima rather than a global optimum result. Due to these limitations various hybrid k-means optimization clustering algorithms are designed.

2. LITERATURE SURVEY

In literature survey the various hybrid k-mean algorithms are discussed like KFFA, KABC, K-Krill herd, KACO, PSO-ACO-K etc.

2.1 Hybrid K-mean Algorithms using Swarm Based Optimization Techniques

Hybrid k-means algorithms are those algorithms that combines k-means algorithm with optimization algorithms. The hybrid k-means algorithms are used to reduce the limitations of k-means algorithm and these are given below.

In KFFA algorithm [5] the k-means clustering algorithm is optimized using firefly algorithm [6]. To find the centroids for specified number of clusters the firefly is used and then to refine the centroids and clusters k-means is applied. In KACO algorithm [7] firstly ACO is applied because cluster quality is based on it. PSO-ACO-K algorithm [8] combines the particle swarm optimization, ant colony optimization and k-means. In KCUCKOO algorithm [9] Cuckoo search leads too much iteration because it randomly selects initial centroids and to overcome this problem they are selected using k-means. In KBAT algorithm [10] the optimization BAT algorithm helps to reduce the local optimal problem of k-means clustering algorithm. In KABC algorithm [11] the k-means is combined with artificial bee colony algorithm for optimization and clusters formed are better than k-means algorithm. In K-Krill herd algorithm [12] the krill herd is used to initialize the centroids for clusters in k-means. Krill herd is used to provide local optimal results. BAT k-medoids [13] clustering algorithm is combined with BAT algorithm to solve the optimization problems of k-medoids algorithm. In KPSO algorithm [14] the results of PSO algorithm is used as the initial seed of the k-means algorithm and k-means algorithm will be applied for refining. The Tabu-KHM algorithm [15] combines the optimization property of tabu search and the local search capability of k-harmonic means algorithm.

2.2 Hybrid K-means algorithms using Bio-inspired Optimization Techniques

In KGA [16] genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection. In KFP algorithm [17] the flower pollination algorithm is used to reduce the disadvantages of k-means local optima and its results are used to select the centroids of clusters in k-means. In KFSS [18] the bio inspired fish school search optimization algorithm is used along with k-means algorithm. K-Means and K-Harmonic with Fish School Search Algorithm [19] provides more optimized results than KFSS. IGSA-KHM algorithm [20] not only helps the KHM clustering escape from local optima but also overcomes the slow convergence speed of the IGSA. CGA [21] is clustering based Genetic Algorithm with polygamy selection and dynamic population control technique. According to CGA the fitness values obtained from chromosomes in each generation were clustered into two non-overlapping clusters. ABCGA means Adaptive Biogeography Clustering based genetic algorithm. In hybrid technique the CGA which means Clustering based Genetic Algorithm is used along with ABPPO which stands for Adaptive biogeography based predator-prey optimization. In the proposed technique the clustering process is similar to k-means algorithm.

3. COMPARATIVE ANALYSIS

In Comparative analysis the various hybrid k-means algorithms are compared based on some output parameters and these comparisons are described below.

3.1 Data set

The data sets used are data sets from the UCI Machine Learning Repository are Wine, Iris, Seed,

Breast Cancer and Liver Disorders data set. Also number of attributes and number of instances in these data sets are described. These data sets are shown in Table 1.

Table 1. Data Sets

S.No	Name	Number of instances	Number of Attributes
1	Wine	179	14
2	Iris	150	4
3	Seed	210	5
4	Breast Cancer	699	10
5	Liver Disorders	345	7

This table shows that the total number of attributes in Wine data set is 14 and number of instances is 179, total number of attributes in Iris data sets is 4 and total number of instances is 150. The total number of attributes in Seed data set is 5 and total number of instances is 210.The total number of attributes in breast cancer data set is 10 and total number of instances is 699. The total number of attributes in Liver disorders data set is 7 and total number of instances is 345.

3.2 Output Parameters

The output parameters are those parameters based upon which the performance of existing clustering algorithm is compared with the new hybrid k-means optimization algorithm. Some output parameters are described below:

3.2.1 TP: True Positive

It is defined as the proportion of positives that are identified correctly. It is also called as Sensitivity. Example: Sick people who are correctly identified as having the condition.

3.2.2 *TN: True Negative*

It is defined as the proportion of negatives that are identified correctly. It is also called as Specificity. Example: Sick people who are correctly identified as not having the condition.

3.2.3 *FP: False Positive*

They are those which are identified incorrectly. Example: Healthy people incorrectly identified as sick.

3.2.4 FN: False Negative

They are those which are incorrectly rejected. Example: Sick people incorrectly identified as healthy.

Some major parameters based upon which the performance of proposed algorithm is evaluated and compared with existing algorithm are described below.

3.2.5 Accuracy

It is defined as only the proportion of true results.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad \text{.....eq. (i)}$$

Here TP = True positive, FP= False positive, TN= True negative, FN= False negative.

3.2.6 *Purity*

Purity is defined as the percent of the total number of objects that were classified correctly. To compute purity each cluster is assigned to the class which is most frequent in a cluster. It describes the cluster quality.

$$\text{Purity} = \frac{1}{N}\sum_k \max_j |w_k \cap c_j| \text{.....eq. (ii)}$$

Where N = number of objects, k = number of clusters, cj is set of classes and wk is the set of clusters.

3.2.7 CPU Time

CPU time means the time taken required by the computer to perform a given set of computations. If CPU time is less than the clustering algorithm is better than algorithms having more Computation time.

3.3 Comparison of K-means, CGA and ABCGA algorithm using Purity

The comparison of CGA algorithm and k-means algorithm with the ABCGA algorithm using Purity when number of clusters is 8 for five data sets that are Wine, Iris, Seed, Breast cancer and Liver Disorders taken from UCI Machine Learning Repository.

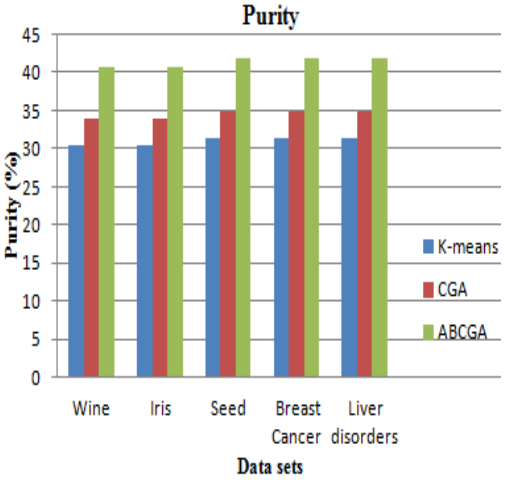

Figure 1: Comparison using Purity

In the above Figure 1 comparison it shows that the ABCGA algorithm purity is high for the all five data sets than CGA and k-means Algorithm for Clustering.

3.4 Comparison of KFFA, KBAT and KFPA algorithm using CPU Time [17]

The comparison of these three KFFA, KBAT and KFPA algorithms are based on the CPU time using two data sets Iris and wine from UCI Machine Learning Repository.

Table 2. Comparison using CPU time

Data set	KFFA	KBAT	KFPA
Iris	8.7	3.2	3.2
Wine	19	3.88	3.87

In the above comparison it shows that the KBAT and KFPA algorithm require less CPU time than KFFA Algorithm for Clustering.

3.5 Comparison of KFSS and KPSO algorithm using Accuracy[19]

The comparison of these two KFSS and KPSO algorithms are based on the Accuracy using two data sets Iris and wine from UCI Machine Learning Repository.

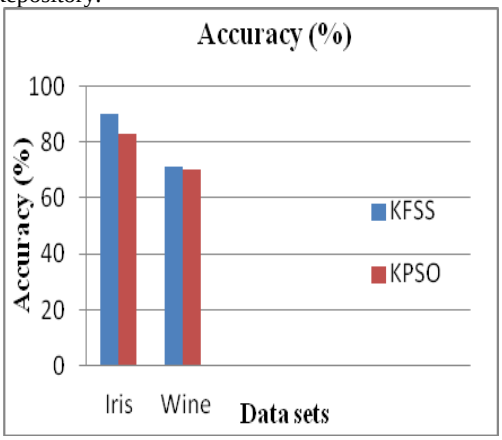

Figure 2: Comparison using Accuracy

In the above comparison it shows that the KFSS algorithm accuracy is high than KPSO Algorithm for Clustering.

The comparative Analysis of various hybrid k-means algorithm is done in the paper using various output parameters. The performance is compared for different data sets that are Iris, Wine, Seed, Breast Cancer and Liver Disorders. The comparison of k-means, ABCGA and CGA is done by using purity output parameter which shows that ABCGA has high purity than other two algorithms for clustering. The KFFA, KBAT, KFPA is compared based on the CPU time whose results shows that KFPA and KBAT requires low CPU time. Another comparison is done using accuracy output parameter for KFSS and KPSO algorithm which shows that the accuracy for KFSS algorithm is high for Iris and wine data sets.

4. CONCLUSION

In this paper, the data clustering, clustering techniques and various hybrid k-mean algorithms are presented. The comparison of the performance of various hybrid k-means optimization algorithms is done. The comparison of CGA and ABCGA algorithm is done through purity which shows that ABCGA algorithm provides better purity than CGA. The KFFA, KBAT and KFPA k-mean hybrid techniques are also compared in this paper. On these the comparative analysis is done on the basis of CPU time and the results show that the KBAT and KFPA requires less CPU time than KFFA. Also the hybrid KFSS and KPSO algorithm are compared based on accuracy and comparison shows that KFSS provides better Accuracy than KPSO. There is scope for improvement in these hybrid k-mean algorithms to handle high dimension data sets.

5. REFERENCES

1. Ming-Syan Chen, Jiawei Han, Ps Yu, “Data Mining: An overview from database perspective” IEEE Transaction on knowledge and data engineering, Vol. 8, Issue 6, pp. 886-883, 1996.

2. Anil .Jan, “Data Clustering: 50 years beyond k-means” Pattern Recognition Letters, Elsevier, Vol. 31, pp. 651-666, 2010.

3. Lior Rokach and Oded Maimon, “Clustering methods” Data mining and Knowledge handbook, pp. 321-352, 2005.

4. J.B. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations” Proceedings of Fifth Berkeley Symposium on Mathematics Statistics and Probability, University of California Press, Vol. 1, pp. 281-297, 1967.

5. S.J Nanda, G. Panda, “A Survey on nature inspired metaheuristic algorithm for partition clustering” Swarm and Evolutionary Computation, Elsevier, Vol. 16, pp. 1-18, 2014.

6. Xin-She Yang, "Firefly algorithms for multimodal optimization" Stochastic Algorithms: Foundations and Applications, SAGA 2009. Lecture Notes in Computer Sciences, Vol.5792, pp.169–178, 2009.

7. K.Aparana and Mydhili K.Nair, “Enhancement of k-means algorithm using ACO as optimization technique on high dimensional data” 2014 international conference on Electronics and Communication Systems (ICECS) IEEE, pp. 1-5, 2014.

8. Taher Niknam and Babak Amiri, “An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis” Applied Soft Computing, Elsevier, vol. 10, pp. 183–197, 2010.

9. Saida Ishak Boushaki, Kamel Nadjet and Omar Bendjeghaba, “ A New Hybrid Algorithm for Document Clustering based on Cuckoo Search and K-means”, Recent advances on Soft Computing and Data Mining SCDM Springer, pp. 59-68, Vol. 287, 2014.

10. Tang Rui, Fong Simon, Yang Xin-She and Deb Sujay, “Integrating nature-inspired optimization algorithms to k-means clustering” 2012 seventh

International Conference on Digital Information Management ICDIM, IEEE, pp. 116-123, 2012.

11. Karaboga K, Dervis D, Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm" Applied Soft Computing, Springer, Vol. 11, pp. 7-652, 2011.

12. Hamed Nikbakht, Hamid Mirvaziri, "A new clustering approach based on K-means and Krill Herd algorithm" 23rd Iranian Conference on Electrical Engineering, IEEE, 2015.

13. Monica Sood and Shilpi Bansal, "K-Medoids Clustering Technique using Bat Algorithm", International Journal of Applied Information Systems, pp. 20-22, 2013.

14. Yucheng Kao, Szu-Yuan Lee, "Combining K-means and particle swarm optimization for dynamic data clustering problems", IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009.

15. Gungor. Z and Unler. A, "k-Harmonic Means Data Clustering with Tabu Search Method" Applied Mathematical Modeling, Vol. 32, pp. 1115-1125, 2008.

16. Md Anisur Rahman, Md Zahidul Islam, "A hybrid clustering technique combining a novel genetic algorithm with K-Means", Knoweldge based systems, Elsevier, 2014.

17. Parul Aggarwal And Shikha Mehta, "Comparative Analysis Of Nature Inspired Algorithm On Data Mining", IEEE International Conference On Research In Computational Intelligence And Communication Networks, 2015.

18. C.J.A. Bastos-Filho, F.B. Lima Neto, A.J.C.C. Lins, A.I.S. Nascimento, M.P. Lima, "A novel search algorithm based on fish school behavior" , in: Proceedings of the 2008 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2646–2651, 2008.

19. Adriane B.S. Serapiao, Guilherme S. Correa, Felipe B. Goncalves, Veronica O. Carvalho, "Combining K-Means and K-Harmonic with Fish School Search Algorithm for data clustering task on graphics processing units" , Applied Soft Computing, Elsevier, Vol. 41, pp. 290-304, 2016.

20. Minghao Yin, Yanmei Hu, Fengqin Yang, Xiangtao Li, Wenxiang Gu, "A novel hybrid K-harmonic means and gravitational search algorithm approach For clustering" Expert Systems with Applications, Elsevier, Vol. 38, pp. 9319-9324, 2011.

21. A.M. Aibinu, H.Bello Salau, Najeeb Arthur Rahman, M.N. Nwohu, C.M. Akachukwu, "A novel Clustering based Genetic Algorithm for route optimization", Engineering Science and Technology an International Journal, Vol. 19, pp. 2022–2034, 2016.