Hybrid Data Cost Setting using K-Means & ACO to Optimize Data Cost

Anita Bishnoi, Mr. Vinod Todwal
Rajasthan College of Engineering for Women, Jaipur, Rajasthan, India

ABSTRACT

In k-means clustering, we are given a set of n data points in d-dimensional space Rd and an integer k and the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. A widespread heuristic for k-means clustering is Lloyd's algorithm. In this paper, we present a simple and efficient implementation of Lloyd's k-means clustering algorithm, which we call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure .We establish the practical efficiency of the filtering algorithm in two ways. First, we present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, we present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

Keywords : ACO, Clusters, K-means, Mahalanobis Distance.

I. INTRODUCTION

K-means clustering algorithm is centered on a partitioning tactic whereby data points are constantly relocated to the nearby centroid and the shape of every single cluster is modified. Though, their performance hinge on a great degree on the initial values of the preliminary centroids randomly produced each time the algorithm is run. It is well-known that K-means just fall into local optima that do not generate the best clustering outcomes. Attaining a globally optimum clustering outcome requires an extensive process in which altogether partitioning prospects are tried out, which is computationally exorbitant. An experiential approach to the problem is to pursuit for global optima in every computational reiteration with the support of an optimization algorithm.

Several bio-inspired optimization procedures have recently occurred with designs rivaling swarm behavior in the midst of groups. Though each of the diverse search agents moves (explore for its track) in its certain way, they swarm together collectively in the way of a collective optimization goal. The bio-inspired optimization algorithms planned to date have added ample courtesy among computer science scientists and researchers. Yet their advantages have been both mathematically calculated and practically applied in several applications, testing of hybrids combining such bio-inspired algorithms with present data mining algorithms rests at an initial phase.

Intuitively, bio-inspired optimization algorithms should overwhelmed the limits of K-means clustering algorithm in discovering globally optimum clusters. The objective of this study is to inspect the fundamental constructs of adding some popular bio-inspired optimization approaches into a K-means clustering algorithm. This addition is substantial, because it will assist as a successful innovator paradigm for the future expansion of fusion bio-inspired data mining algorithms.

CSEIT1831178 | Received : 15 Jan 2018 | Accepted : 05 Feb 2018 | January-February-2018 [(3) 1 : 755-760]

Even if some scholars have made an effort to associate K-means algorithms with bio-inspired algorithms, their efforts have been limited to nearly the same form of cluster travels, such as the genetic Algorithm, Firefly, Ant Colony optimization and Particle Swarm Optimization algorithms. Moreover just a small number of bio-inspired optimization algorithms combined with K-means are obtainable in the previous papers. When the number of bio inspired optimization techniques are positioned simultaneously, with an aim of enlightening K-means clustering, some common prototypes can be observed. In our work we try to present a outline of a broad proposal of improved K-means clustering with bio-inspired optimization technique. To authenticate the effectiveness of the framework, trials that are founded on synthetic and real life datasets are implemented.

II. LITERATURE SURVEY

Amira Boukhdhir, Oussama Lachiheb, Mohamed Salah Gouider[1] projected algorithm an enhanced KMeans using Map Reduce design for extensive dataset. The algorithm receipts less performance time as compared to traditional KMeans, PKMeans and Fast KMeans. It removes the outlier from numerical datasets also Map Reduce procedure used to choose initial centroids and formation of the clusters. But it has boundaries like the value of amounts of centroids requisite as input by user. It works on arithmetical datasets only. Also the amount of clusters are not determined inevitably.

Agustin Blas et al.[2] defined the performance of the grouping Genetic algorithm in clustering, initiated with projected encoding, and altered modification of crossover and mutation process and also started the native search contain with the island model for enhance the performance of the difficult situation.

The real data sets were used and combined the outputs with the standard formulas such as DBSCAN and K-means, and gaining the exceptional outcomes in planned grouping based procedure the evolutionary method such as Genetic algorithm. The results of the algorithm was calculted by using the different fitness function.

Tzung-Pei-Hong et al.[3] describe the performance of the Genetic algorithm based aspect clustering method were enhanced based on the grouping Genetic algorithm. The chromosome illustration, Genetic procedures, and fitness utility defined in grouping Genetic algorithm to resolve the clustering problem. The output of grouping Genetic algorithm based clustering algorithm improved the merging speed and fitness value of the clustering problem. In addition the algorithm also deal with the problem of lost values. The other optimization algorithms are used to resolve the problem in attribute grouping.

Daniel Gomes Ferrari et al. [4] discussed a new mode to signify the clustering problems molding on the similarity between the objects and the process to associate the interior indication for ranking algorithms designed on the performance of the problem. The practical results defined the viability of meta learning systems for an unspecified mode to the clustering algorithm selection problem. This technique presents the better result from the distance based data set over the attribute based approach.

Kunnuri Lahari et al. [5] reduce the local minima using progressive and improvement based methods like Genetic algorithm and explaining gaining knowledge based optimization. The general data sets used, and the experimental outcomes are compared with the Genetic algorithm and explaining gaining knowledge based optimization clustering with k-means algorithm. The result of the progressive clustering method compared with some existing clustering process.

Rajashree Dash et al.[6] defined a comparative analysis of K-means and Genetic algorithm rely on clustering.

Arun Prabha et al.[7]enhanced the cluster quality from K-means clustering by applying a Genetic algorithm. huge scale clustering problems in data mining also noticed by this process. The best outcomes are acquired by applying this process.

Anusha et al. [8] gives an enhanced K-means Genetic algorithm for optimization of cluster. It overcomes the drawback of local optima with suitable dataset and also the algorithm fails in execution time. It is concluded that the technique created more than the 90% correctness for actual dataset. It is also adopted a vicinity knowledge strategy for optimizing multi objective problems. This algorithm apply k means Genetic algorithm to find the effectiveness of the clusters. It is found that the algorithm could produce smallest index value for the largest datasets.

Xiaoli Cui and al [9] defined an algorithm i. e. an enhanced k-means. This algorithm is applied on only illustrative points in its place of the whole dataset, applying a sampling process. The outcome of this the I/O cost and the network cost decreased because of Parallel K-means. Practical outcomes defines that the algorithm is effective and it has better results as compared to k-means but, there is no high exactness.
Yugal Kumar, G. Sahoo [10]defined K-Means initialization problems. The K-Means initialization problem of algorithm is given by two types; first one is the number of clusters mandatory for clustering and second one is to find the way to initialize starting centers for clusters of K-Means algorithm. In this paper the solution for of the initialization problem of initial cluster centers is found. To find that solution a binary search initialization process is used to initialize the starting cluster points.

Huang Xiuchang, SU Wei [11] defined the problem of user behavior design analysis, which has the insensitivity of arithmetical value, odd three-dimensional and temporal dissemination characteristics strong noise. The old-style clustering algorithm not works correctly. It analyses the previous clustering process, trajectory analysis process, and behavior design analysis process, and mix the clustering algorithm into the trajectory analysis. After updating the old-style K-MEANS clustering algorithm, the new enhanced algorithm created which is applicable to resolve the problem of user behavior design analysis compared with previous clustering process on the basis of the results of the simulation data and real data, the outcome shows that the enhanced algorithm more applicable to solve the trajectory pattern of user behavior problems.

III. PROPOSED WORK

In this section, the base of k means has been explained and with the implemented changes and work, K-means method sectionalize the data and the forms the cluster depending upon the number of cluster user wants to form and also over the length of the data sets, however the condition only suggests that the number of clusters should not be greater than the length of the database.

In the presented paper the K-means methodology has been overturned and features like cluster calculation and density fixation are added, also the distance formula has been changed as due to the fact that the previously used Euclidean function or the squared distance formula is slow in process and hence took longer time.

To analyze the big data and the nature of the certain data related to any particular field, there are certain methodologies to perform such data analysis and also help in predicting the upcoming results or the outcomes of the certain data based on their characteristics and the environment.
Here, we are giving the K-Means clustering method with the bio inspired algorithm.

k-means clustering is a procedure of vector quantization, previously from signal processing, that is prevalent for cluster analysis in data mining. k-means clustering purposes to partition n data elements into k clusters in which each data elements

fit in the cluster with the nearby mean, helping as a pattern of the cluster. This outcomes in a subdividing of the data space into Voronoi cells.
The problem is computationally tough (NP-hard); though, there are effective experiential algorithms that are usually employed and congregate rapidly to a local optimum. These are generally alike to the expectation-maximization algorithm for combinations of Gaussian distributions by an iterative enhancement approach employed thru both algorithms. Moreover, they both usage cluster centers to model the data; though, k-means clustering inclines to search clusters of comparable spatial degree, whereas the expectation-maximization mechanism permits clusters to have dissimilar shapes.

Bio Inspired algorithms are those algorithm which carry forwards the traits of the living bodies or sometimes naturally non-living objects. Algorithms which can be used in this paper alongside the k-means method are genetic algorithm, particle swarm optimization or ant colony optimization.

Algorithm 3 Modified K-means

```
1. Collect the database from the server of 6instruments.com
2. Collected data is the available online streaming data of the big giants like AMAZON, Twitter, Facebook, Tumblr, Linkedin and Pininterest.
3. Calculate the length of the collected database, n = len(database)
4. Define the number of cluster using nCr function, where n is the length of the database and r is the maximum density of the cluster.
5. Apply the k-means method.
   5.1 Select centroids
   5.2 Calculate distance between the points using the getdistance formula, which includes the distance formula of Chebyshev or mahalanobis or manhattan.
   5.3 Create clusters.
```

Ant Colony Optimization: In the natural world, ants of some species (initially) wander randomly, and upon finding food return to their colony while laying down pheromone trails. If other ants discover such a path, they are to be expected not to keep travelling at random, but in its place to follow the trail, returning and strengthening it if they ultimately find food.

Over time, though, the pheromone trajectory starts to evaporate, thus decreasing its attractive strength. The additional time it receipts for an ant to travel down the track and back again, the additional time the pheromones have to evaporate. A short track, by evaluation, gets streamed over more frequently, and therefore the pheromone concentration becomes higher on shorter tracks than longer ones. Pheromone desertion also has the benefit of avoiding the merging to a locally optimal solution. If there were no desertion at all, the tracks chosen by the first ants would tend to be extremely attractive to the following ones. In that circumstance, the exploration of the result space would be constrained. The effect of pheromone evaporation in real ant organizations is vague, but it is very significant in artificial systems.

The whole outcome is that when one ant discovers a good (i.e. small) track from the colony to a food source, other ants are much prospective to follow that track, and optimistic feedback ultimately leads to all the ants following a single track. The notion of the ant colony algorithm is to mimic this performance with "simulated ants" walking around the graph on behalf of the problem to solve.

Here in this paper the two modifications that have been included are the distance formula optimization and the other one is the path finder, ACO is used to detect the path of the cluster to optimize the data threading, which reduces the time for the K-Means algorithm. The distance formula used for determining the shortest between the centroid and the other nodes and between the nodes for the ACO algorithm is mentioned in the equation below.

Chebyshev Distance formula:

IV. RESULTS

The proposed algorithm was simulated on a cluster of 2 machines having the configuration as stated in Table 1. All the nodes were connected by a 100 Mbps Ethernet switch with Windows 7 Ultimate Edition operating system. The Python 3.4 version, Pycharm 2017.1 version was installed.

Table 1

Features	**Machine 1**	**Machine 2**
Processor	Core i5 1st generation	Core i5 3rd Generation
RAM	6GB	8GB
Graphics	1GB ATI Radeon	2GB NVidia
Hard Disk	500GB	1TB

Experimental results are shown in the form of the graphs generated during the process run over the Pyhton 3.4.4.

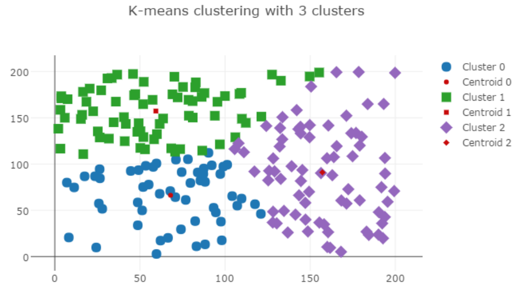


Figure 1. Dataset Length: 200, K-means Hybrid Clusters: 3, Assumed Density: 112

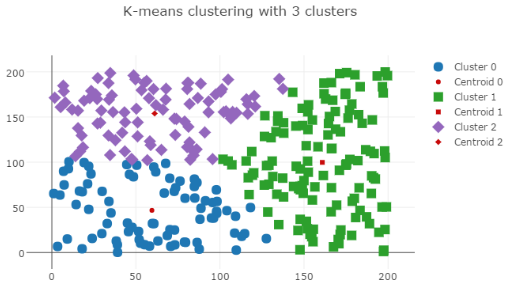


Figure 2. Dataset Length: 200, K-means Hybrid Clusters: 3, Assumed Density: 100

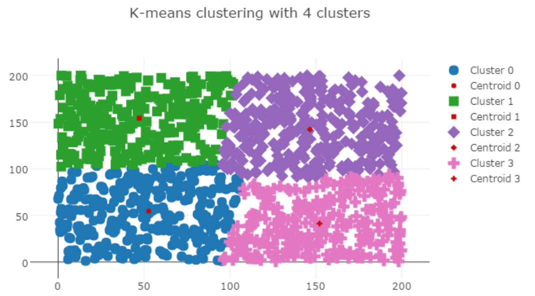


Figure 3. Dataset Length: Amazon Book Views 1295, K-means Hybrid Clusters: 4, Assumed Density: 323.75

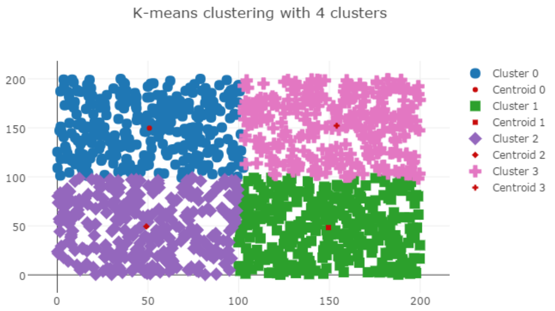
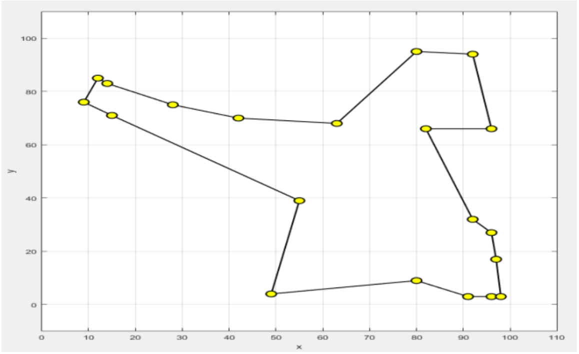


Figure 4. Dataset Length: 1295, K-means Hybrid Clusters

The time noted on the simple K-Means methodology typically appears to be around 4 minutes and with same database it is noted out to be 37.5 seconds.



Results On Different Iterations :

Iteration 1: Best Cost = 632.5771

Iteration 2: Best Cost = 530.2894

Iteration 4: Best Cost = 476.6628

Iteration 5: Best Cost = 443.0602

Iteration 9: Best Cost = 399.9723

Iteration 25: Best Cost = 381.2995

Iteration 37: Best Cost = 370.616

```
Iteration 103: Best Cost = 362.038
Iteration 2999: Best Cost = 362.038
Iteration 3000: Best Cost = 362.038
Elapsed time is 111.204726 seconds.
```

V. CONCLUSION

The work presented here excels on various parameters when simulated and hence the techniques implemented here in this paper to achieve the target is commendable and easy to implement with the Real Time systems, however the inclusion of neural networks and Artificial Intelligence in order to determine the number of clusters will be appreciable.

VI. REFERENCES

[1]. AmiraBoukhdhir Oussama Lachiheb, Mohamed Sala Gouider. "An improved Map Reduce Design of Kmeans for clustering very large datasets", IEEE transaction.

[2]. Huang Xiuchang , SU Wei ,"An Improved K-means Clustering Algorithm" ,JOURNAL OF NETWORKS, VOL. 9, NO. 1, JANUARY 2014

[3]. Yugal Kumar and G. Sahoo, "A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm", International Journal of Advanced Science and Technology Vol.62, (2014).

[4]. Honga Tzung-Pei, Chun-Hao Chenc, Feng-Shih Lin, "Usinggroup genetic algorithm to improve performance of attribute clustering," Elsevier, pp.1-8, 2015.

[5]. Danial Gomes Ferrari, Leandro Numes de Castro, " Clustering algorithm selection by meta-learning systems: A new distance based problems characterization and ranking combination methods," Elsevier, pp.181-194, 2015.

[6]. Rajashree Dash and Rasmita Dash, "Comparative analysis of K means and Genetic algorithm based data clustering," International Journal of Advanced Computer and Mathematical Sciences, pp.257-265, 2012.

[7]. Edvin Aldana-Bobadhilla, Angel Kuri-Morales, "A Clustering based method on the maximum entropy principle," Entropy Article, pp. 151-180, 2015.

[8]. Kannuri Lahari, M. Ramakrishna Murty, and Suresh C. Satapathy, "Prediction based clustering using genetic algorithm and Learning Based Optimization Performance Analysis," Advances in Intelligent Systems and Computing," pp. 338, 2015.

[9]. Rahila H. Sheikh, M. M.Raghuwanshi, Anil N. Jaiswal, "Genetic algorithm based clustering: A Survey," IEEE, pp.314-319, 2008. C. J. Kaufman, Rocky Mountain Research Laboratories, Boulder, CO, private communication, 2004.

[10]. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Jpn., vol. 2, pp. 740-741, August 1987 [Dig. 9th Annual Conf. Magn. Jpn., p. 301, 1982].

[11]. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.