Murala Madhusudhanarao * et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.6, Issue No.3, April - May 2018, 8027-8029.

Stretched Reliable Metaheuristics for Enhancing K-means

MURALA MADHUSUDHANARAO
PG Scholar, Department of CSE, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh.

Dr. G.V.S.N.V PRASAD
Professor & Dean Academic Affairs, Department of CSE, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh.

Abstract: **Cluster analysis is one of the primary data analysis methods and k-means is one of the most well known popular clustering algorithms. The k-means algorithm is one of the frequently used clustering method in data mining, due to its performance in clustering massive data sets. The final clustering result of the kmeans clustering algorithm greatly depends upon the correctness of the initial centroids, which are selected randomly. The original k-means algorithm converges to local minimum, not the global optimum. Many improvements were already proposed to improve the performance of the k-means, but most of these require additional inputs like threshold values for the number of data points in a set. In this paper a new method is proposed for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity. According to our experimental results, the proposed algorithm has the more accuracy with less computational time comparatively original k-means clustering algorithm.**

Keywords: **Clustering; Data Mining; Data partitioning; Initial cluster centers; K-means clustering algorithm. Cluster analysis.**

1. INTRODUCTION:

Clustering is the process of organizing data objects into a set of disjoint classes called clusters. Clustering is an example of unsupervised classification. Classification refers to a procedure that assigns data objects to a set of classes. Unsupervised means that clustering does not depends on predefined classes and training examples while classifying the data objects.Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. Therefore, a cluster is a collection of objects that are similar among themselves and dissimilar to the objects belonging to other clusters. Clustering is an crucial area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics, etc. Cluster analysis is a one of the primary data analysis tool in the data mining. Clustering algorithms are mainly divided into two categories: Hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the data set into desired number of sets in a single step [9]. Numerous methods have been proposed to solve clustering problem. One of the most popular clustering method is kmeans clustering algorithm developed by Mac Queen in 1967. The easiness of k-means clustering algorithm made this algorithm used in several fields. The k-means clustering algorithm is a partitioning clustering method that separates data into k groups [1], [2], [4], [5], [7], [9]. The k-means clustering algorithm is more prominent since its intelligence to cluster massive data rapidly and efficiently. However, kmeans algorithm is highly precarious in initial cluster centers. Because of the initial cluster centers produced arbitrarily, kmeans algorithm does not promise to produce the peculiar clustering results. Efficiency of the original k-means algorithm heavily rely on the initial centroids [2], [5]. Initial centroids also have an influence on the number of iterations required while running the original k-means algorithm. The computational complexity of the original k-means algorithm is very high, specifically for massive data sets [2]. Various methods have been proposed in the literature to enhance the accuracy and efficiency of the k-means clustering algorithm. This paper presents an enhanced method for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity

2. EXISTING SYSTEM:

As the volume of data generated increases significantly, new challenges arise daily to understand and explore such data. Working with large datasets typically aims to extract useful knowledge from them. When extracting knowledge from data, one often needs to apply clustering techniques, either as a preprocessing step for or as the final goal of the data analysis task. Data clustering aims to split the dataset into a finite number of categories according to the similarity or interrelationships among the data objects. It is an unsupervised technique, i.e., no class labels are provided. The applicability of clustering algorithms includes areas such as image processing, document categorization, and bioinformatics etc. Among the clustering algorithms, k-means is considered one of the ten most influential algorithms in data mining, mainly due to its simplicity, scalability, and for being easy to adapt to different application scenarios and domains. The difficulty to adapt

many classic clustering's algorithms for parallel and distributed models is mainly due to the fact that most algorithms were not originally designed to work under big data paradigm. So a better system is required to support an updated clustering paradigm on continuous evolving data (Big).

3. THE K-MEANS ALGORITHM

One of the most popular clustering method is k-means clustering algorithm. It generates k points as initial centroids arbitrarily, where k is a user specified parameter. Each point is then assigned to the cluster with the closest centroid. Then the centroid of each cluster is updated by taking the mean of the data points of each cluster. Some data points may move from one cluster to other cluster. Again we calculate new centroids and assign the data points to the suitable clusters. We repeat the assignment and update the centroids, untill convergence criteria is met i.e., no point changes clusters, or equivalently, until the centroids remain the same. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids. Pseudocode for the k-means clustering algorithm is described in Algorithm

1. The Euclidean distance between two multi-dimensional data points *X = (x1, x2, x3... xm)* and *Y = (y1, y2, y3... ym)* is described as follows:

$$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+..+(x_m-y_m)^2}$$

Algorithm 1: The k-means clustering algorithm [2]

Require: *D = {d1, d2, d3, ..., di, ..., dn }* // Set of n data points.

k // Number of desired clusters

Ensure: A set of *k* clusters.

Steps:

1. Arbitrarily choose *k* data points from *D* as initial centroids;

2. **Repeat**

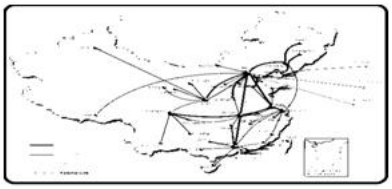
Assign each point *d*i to the cluster which has the closest centroid;

Calculate the new mean for each cluster;

Until convergence criteria is met.

4. RELATED WORK

The original k-means algorithm is very impressionable to the initial starting points. So, it is quite crucial for k-means to have refine initial cluster centers. Several methods have been proposed in the literature for finding the better initial centroids. And some methods were proposed to improve both the accuracy and efficiency of the k-means clustering algorithm. In this paper, some of the more recent proposals are reviewed [1-5], [8].Proposed an enhanced method for assigning data points to the suitable clusters. In the original kmeans algorithm in each iteration the distance is calculated between each data element to all centroids and the required computational time of this algorithm is depends on the number of data elements, number of clusters and number of iterations, so it is computationally expensive.The required computational time is reduced when assigning the data elements to the appropriate clusters. But in this method the initial centroids are selected randomly. So this method is very sensitive to the initial starting points and it does not promise to produce the unique clustering results.



System Architecture

5. DISTRIBUTED K-MEANS

Basic k-means as the disadvantage of not supporting the Big data concept, so moving to proposed system of distributed k-means. In this Distributed k-means the clustering concept will be there, so easily the data can be splited into the records. The clustering results will be easily stored as a graphs.

6. ENHANCEMENT

1. Prior approaches handled big data to reduce clusters from distributed architecture very well.

2. Although it counters node failures, it cannot filter out real time data anomalies.

3. An often encountered problem with respect to distributed architecture is the problem ofinsufficient/corrupted/unavailable data points which detriments the formation of clusters to exception recoveries.

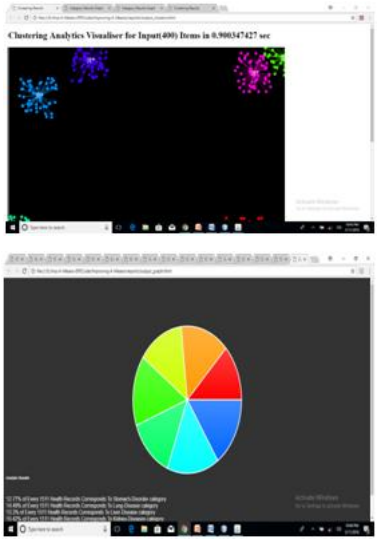
4. Although the necessary cluster points are generated the complexities involved in ignoring the missing tuples during run time is a significant constraint that needs to be handled.

5. So we propose a Bloom Filter oriented early pruning algorithm sorted positional indexes to validate data points during run time with a small pre-constructed data-structures to reduce I/O cost significantly.

6. Thus, the intuitive idea of pruning is to omit the tuples that are not of any help as much as possible. In this way, the CPU cost and I/O cost can be reduced significantly.

7. Pruning operations on the candidate positional indexes, and its abstraction.

Algorithm 1. EarlyPruning(*j*, pil, pit).
// *testInBF(bf, pi) checks whether pi belongs to S on which*
// *bloom filter bf is constructed, true is in, false is not.*
// *j is the index for sorted-positional-index-list*
// *pil is the positional index in sorted-positional-index-list*
// *pit is the candidate positional index for T*
// *return: true - pit can be pruned, false - pit cannot be pruned*
1: int $index_j = \lceil log_2 PC_j \rceil // (PC_j = max\{C_k^{pru}[j] \ (1 \le k \le m)\})$
2: **if** $pil \le RC_j$ **then**
3: return false $//(RC_j = min\{C_k^{pru}[j] \ (1 \le k \le m)\})$
4: **end if**
5: **for** $k = 1$ to m **do**
6: **if** j == k **then**
7: continue
8: **end if**
9: boolean inflag = testInBF($EGBFT_k(index_j)$, pit)
10: **if** (inflag) **then**
11: return false
12: **end if**
13: **end for**
14: return true



Results obtained through these implementations validates our filtering efficiency.

7. CONCLUSION

One of the most popular clustering algorithm is k-means clustering algorithm, but in this method the quality of the final clusters rely heavily on the initial centroids, which are selected randomly. Moreover, the k-means algorithm is computationally very expensive also. The proposed algorithm is found to be more accurate and efficient compared to the original k-means algorithm. This proposed method finding the better initial centroids and provides an efficient way of assigning the data points to the suitable clusters. This method ensures the total mechanism of clustering in O(nlogn) time without loss the correctness of clusters. This approach does not require any additional inputs like threshold values. The proposed algorithm produces the more accurate unique clustering results. The value of k, desired number of clusters is still required to be given as an input to the proposed algorithm. Automating the determination of the value of k is suggested as a future work.

REFERENCES:

[1] M. Xu, Y. Shang, D. Li, and X. Wang, “Greening data center networkswith throughput-guaranteed power-aware routing,” Comput.Netw., vol. 57, no. 15, pp. 2880–2899, 2013.

[2] R. Kubo, J. Kani, H. Ujikawa, T. Sakamoto, Y. Fujimoto, N.Yoshimoto, and H. Hadama, “Study and demonstration of sleepand adaptive link rate control mechanisms for energy efficient10G-EPON,” IEEE/OSA J. Opt. Commun. Netw., vol. 2, no. 9,pp. 716–729, Sep. 2010.

[3] J. L. Sobrinho, “Algebra and algorithms for QoS path computationand hop-by-hop routing in the internet,” IEEE/ACM Trans. Netw.,vol. 10, no. 2, pp. 541–550, Aug. 2002.

[4] Y. M. Kim, E. J. Lee, H. S. Park, J.-K. Choi, and H.-S. Park, “Antcolony based self-adaptive energy saving routing for energy efficientInternet,” Comput. Netw., vol. 56, no. 10, pp. 2343–2354, 2012.

[5] J. Wang and K. Nahrstedt, “Hop-by-hop routing algorithms forpremium-class traffic in DiffServ networks,” ACM SIGCOMMComput. Commun. Rev., vol. 32, pp. 73–88, 2002.

[6] Yuan Yang, Student Member, IEEE, Mingwei Xu, Member, IEEE,Dan Wang, Member, IEEE, and Suogang Li, “A Hop-by-Hop Routing Mechanismfor Green Internet”, ieee transactions on parallel and distributed systems, vol. 27, no. 1, january 2016.

[7] O. Heckmann, M. Piringer, J. Schmitt, and R. Steinmetz, “Generating realistic ISP-level network topologies,” IEEE Commun. Lett., vol. 7, no. 6, pp. 335–336, Jul. 2003.

