Hybrid clustering algorithm ‘KCu’ for combining the features of K-means and CURE Algorithm for efficient outliers handling

B. Renuka Devi1*, S. Pallam Setty2

1Department of CSE, Vignan’s Nirula Institute of Technology & Science for Women, Guntur 522005, Andhra Pradesh, India
2Department of CS & SE, College of Engineering, Andhra University, Andhra Pradesh 530003, India

Corresponding Author Email: dr.b.renukacse@gmail.com

https://doi.org/10.18280/ama_b.610204

Received: 26 April 2018
Accepted: 2 June 2018

Keywords: big data, clustering, partitioning, hierarchical k-means, CURE hybrid algorithm

ABSTRACT

In the ongoing situation, the volume of information expands step by step. By the year 2020 the volume of Big Data would reach up to 40zb according to International Data Corporation (IDC). Big Data has turned out to be prevalent for handling, putting away and overseeing huge volumes of information. The grouping of datasets has turned into a testing issue in the field of Big Data examination; however, there are entanglements for applying conventional bunching calculations to huge information because of expanding the volume of information step by step. In this manuscript a new hybrid clustering algorithm, namely KCu to combine the features of both K-Means and CURE clustering algorithms is proposed. The proposed algorithm first applies k-means on data set and then applies CURE on resultant clusters from k-means. We experimented KCu and we show that, when compared to k-means and Cure. Which gives accurate results because of CURE? CURE can handle outliers and it gives non spherical shapes it is the disadvantage of other clustering algorithm.

1. INTRODUCTION

As of late the term enormous information has turned out to be extremely prevalent word in each field. The volume of information builds day by day. As per International Data company (IDC), the volume of Big Data would reach up to 40zb, by the year 2020. Big data comes from three major sources, which are machine data this data comes from industrial equipment, Social data this data comes from Facebook likes, comments, sharing, Twitter tweets r-tweets and YouTube views. Which are growing in an unconceivable range. Public web is another source of social data, this can increase the volume of big data. Transactional data this can be data from payment orders, delivery records, storage records and invoices. There is boundless amount of data that has been generated by the systems using sensors by data accession techniques. Big data refers to complex and large datasets which cannot be processed using traditional databases.

The data can be statistical, events, correlations, and hypothetical. Veracity refers Uncertainty due to data inaccurate, inclination, and abnormality. The quality of the data being captured depends on the veracity of the source data. Big data has very low density which means one single observation does not have any significance of its own. With the massive amount of data being generated by people and organizations today, big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Presently we are having four categories of analytics those are predictive analytics, diagnostic analytics, descriptive analytics and prescriptive analytics. Particularly, descriptive analytics place major role in getting a pattern from unstructured data. In this paper we are proposing a novel mechanism to analyse unstructured data. In our mechanism we are proposing a hybrid algorithm based on CURE and K-Means clustering algorithms.

2. RELATED WORK

Rampdas Raghavan et al. [1] they proposed architecture for k-means clustering with quick and scalable dependent equivalent processing architecture, consider a profound research on K-means algorithm calculation of streamlining. They were set a forward the primary chose starting bunching focus of K-means calculation, toward this end.

Chuan Liu [2], et al. they proposed a profound research on K-implies calculation of advancement. They set forward the main chose introductory grouping focal point of K-means calculation, toward this end, a novel half and half calculation in view of K-means algorithm and Hybrid Rice Optimization Algorithm they were proposed to quickly locate the ideal bunch focuses and abstain from getting into nearby ideal.

Caiquan Xiong [3] et al. proposed a k-means algorithm, traditional algorithm has drawback to choose initial centroids. They introduce an enhanced k-means algorithm, this algorithm initially figures the density of every data object in dataset and after that dole out which dataset is near to centroid.

Jehyun Karimov [4] et al. proposed a novel hybrid evolutionary model. This model aims to choosing good candidates for initial centroids, which uses the meta-heuristic methods to choose the centroids in k means clustering algorithm, this model integrates the PSO, SS and SA, but this model has drawback we can’t give the large datasets.

Jongkyu Han [5] et al. proposed a quick k-means strategy in light of factual bootstrapping method. They were proposed strategy accomplishes approximately 100 times speedup and
comparative precision contrasted with Lloyd calculation it is the well-known k-means calculation in modern field.

Anupama Chadhaan Suresh Kumar [6] they proposed an enhanced k-means algorithm. k-means is division based clustering algorithm. This algorithm developed a density based detection methods. K-means algorithm is very efficiently but it has deficiencies which are the number clusters needs, centroid selection and noisy data points. Here they were use noise data filter to advanced k-means algorithm.

Shi Na [8], et al. with a specific end goal to fathom this inquiry, requiring a basic data structure to store some data in each iteration, which is to be utilized as a part of the following interation. The enhanced strategy abstains from registering the separation of every datum protest the group focuses repeatedly, sparing the running time.


3. CLUSTERING TECHNIQUES

A. K-Means Clustering Algorithm

K-Means is one of the most popular and commonly used algorithms. Proposed by McQUEenin 1976. K-Means is a partition based clustering algorithm, and it is a mathematical, non-deterministic, unsubstantiated, and iterative method. This algorithm is proved a very efficient way then it can produce good clustering results. It is probing data analysis technique that is it explore the complete data set. K-means implements non-hierarchical method of grouping objects together. But it will take the data set as coming and then it will group them. The informational index is isolated into K bunches in light of least Euclidean distance. At the moment that each one of the information objects are consigned to a few bunches, now initial step is done and primary alignment is done. Second step is to recalculate the centroid of the group. Thusly k centroids are iteratively changing their situations in each move until there is no variety in centroid esteems. The computational unpredictability of this calculation is O(mkt), where t is the quantity of cycles. This calculation considered as direct, and K-Means calculation can ready to treat immense datasets. And furthermore this calculation is exceptionally straightforward and simple to execute this is the reason this calculation has been utilized generally. K-Means shows the requirement to non-numerical information.

3.2 CURE (Clustering Using REpresentatives) algorithm

CURE is a hierarchical clustering algorithm. Proposed by Guha, Rastogi and shin in 1998. CURE implements between centroid and all points techniques. It usually within Euclidean method to find out centroid and how points will be placed in cluster. It has a pre define representative points. It works well with outliers which was disadvantage of the other clustering approach and it shrinks the cluster with the factor of ‘a’ that is Euclidean distance that we calculate with that it shrinks the cluster.

![CURE architecture](image)

**Figure 1. CURE architecture**

Random sampling: The random samples are considered as input of algorithm. And it is fitted in main memory. Random samples are generated very fast

Partitioning sample: Partitioning helps to speed up the CURE algorithm. It will Partition the data points into different partitions (n/p). The upside of dividing the input is to decrease the execution time. Each n/p group of points fit in the main memory for increasing performance of partial clustering.

Handling outliers: Random sampling filter out the bulk of outliers. Outliers, because of their bigger separation from the points tend to converge with different points, and develop slower. Number of outliers are less than then clusters. The
bunches which are developing gradually are identified and also very s clusters.

**Labelling data on disk:** The process of sampling initial data sets, the data points are assigned. Each cluster is created using the representative’s points. Merging the close representative’s points. And data return on to the disk.

**Input:** k number of clusters, s objects of database D  
**Output:** A set of k clusters

1. Draw random s samples from initial data set.
2. Divide s samples into p partitions then each contain equal size s/p data objects.
3. Partially cluster each partition until final number of cluster created reduces to s/(p*q) with q>1.
4. Eliminate outliers by using random sample.
5. Cluster partially clusters.
6. label clustering by Representatives and return to the disk.

Contracting the scattered focuses toward the mean disposes of surface anomalies and lessening the endeavours of exceptions. The clusters with the closest pair of representatives are clustered and merged at each step in CURE. The process of choosing c becomes slow as the number of pints in cluster increases. The new points are scattered to the boundaries. CURE can detect clusters of non-spherical shape, with variation of size the representative’s points for each cluster. Good execution time with large database and sets using random sampling and partition methods. Works well with outliers, which are detected and merged or eliminated.

4. **HYBRID CLUSTERING METHOD**

A new clustering algorithm called KMC is proposed, which is based on the K-Means and CURE clustering algorithms.

**Figure 2. Architecture for hybrid clustering algorithm**

This hybrid KCu clustering algorithm has two phases first is to apply k means clustering algorithm on initial data sets. This algorithm has two steps first step is to choose the k the number of clusters that we need. Next we have to choose random data objects from dataset as initial centroids. And then calculate distance between centroid and data objects using Euclidean distance, then after assign each data object to the nearest centroids based on minimum Euclidean distance.

4.1 **Hybrid clustering algorithm**

**Input:** k // Desired number of clusters  
**D= {d1, d2,……..,dn} // set of data objects**  
**Output:** K={c1, c2,………..ck} // Set of k clusters  

1. Assign initial values for centroid point m1, m2, ..., mk // k seeds
2. Repeat
   1. Calculate Euclidean distance from m_k to c_k 
   2. Assign each c_data object to the nearest centroids m_k based on minimum Euclidean distance
3. Calculate new centroid for each cluster; until the centroids do not change any more.
4. Apply CURE on the resultant clusters K={c1, c2, ……..ck}
5. Draw random sample s
6. Partition sample to p partitions with size s/p.
7. Partially cluster each partition until the final number of cluster created reduces to s/(p*q) with q>1.
8. Eliminate outliers by using random sample
9. Cluster partially clusters
10. Cluster partially clusters
11. label data in disk

At the point when every one of the data objects are relegated to some clusters, now first step is done and primary alignment is finished. Second step is to recalculate the centred of the cluster. Along these lines k-centroids are iteratively changes their positions in every move until there is no variation in centroid values. Second phase is to apply CURE on resultant clusters. In this algorithm first we have to draw random sample and then partition the random samples. Often that we have to partially cluster the partition sample. Now we need to eliminate outliers, now again we have to cluster partial clusters then return the data on to disk finally we got efficient clusters. Using this re-clustering we can eliminate outliers and also we get both spherical and non-spherical shapes with unique shapes.

5. **RESULTS**

When the proposed algorithm is used performance accuracy based on the size of the dataset used is illustrated in figure 3.
6. CONCLUSION

In this paper we introduce a new hybrid clustering algorithm which is based on K-Means and CURE clustering algorithms. Here we done a re-clustering, which means first we have to apply k-means clustering algorithm on initial large data sets and the we apply CURE algorithm on resultant clusters from k-means then we get accurate results, by we are applying CURE ,we get non-spherical shapes which is disadvantage of other clustering algorithms and also it can handle outliers efficiently than k-means, then we get accurate results than other clustering algorithms.

REFERENCES