ISSN:0975-9646

Enhanced K-Means Clustering Algorithm to Reduce Time Complexity for Numeric Values

Bangoria Bhoomi M.
PG Student [C.E.], Noble Engineering College, Junagadh, Gujarat

Abstract -Data mining is the process of using technology to identify patterns and prospects from large amount of information. In Data Mining, Clustering is an important research topic and wide range of unverified classification application. Clustering is technique which divides a data into meaningful groups. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. In this paper, try to improve the performance using some changes in already existing algorithm.

Keywords:Clustering, data mining

I. INTRODUCTION

Fast rescue of the related information from the databases has always been a major issue. Different techniques have been developed for this purpose, one of them is Data Clustering.Clustering is a technique which divides data objects into groups based on the information establish in data that illustrates the objects and relationships between them, their mark values which can be used in many applications, such as knowledge discovery, vector quantization, pattern recognition, data mining, data dredging and etc. [2]

A categorization of major clustering methods: [1]

A. Partitioning methods

In partitioned clustering, the algorithms typically determine all clusters at once, it divides the set of data objects into non-overlapping clusters, and each data object is in exactly one cluster.

B. Hierarchical methods

It creats a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds.

C. Density-based methods

Density-based methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. i.e. A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density (DBSCAN.)

D. Grid-based methods

Grid-based methods quantize the object space into a finite number of cells that form a grid structure.the main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. i.e. STatistical INformation Grid (STING)

E. Model-based methods

Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. i.e. Expectation-Maximization (EM)

F. High-dimensional data clustering methods

It is a particularly important task in cluster analysis because many applications require the analysis of objects containing a large number of features or dimensions. i.e. CLustering In QUEst (CLIQUE) : A Dimension-Growth Subspace Clustering Method

G. Constraint-based clustering methods

It is a clustering approach that performs clustering by incorporation of user-specified or application-oriented constraints.

II. K-MEANS CLUSTERING ALGORITHM

K-means is a data mining algorithm which performs clustering. As mentioned previously, clustering is dividing a dataset into a number of groups such that similar items fall into same groups [1]. Clustering uses unsupervised learning technique which means that result clusters are not known before the execution of clustering algorithm unlike the case in classification. Some clustering algorithms takes the number of desired clusters as input while some others decide the number of result clusters themselves.

K-means algorithm uses an iterative process in order to cluster database. It takes the number of desired clusters and the initial means as inputs and produces final means as output. Mentioned first and last means are the means of clusters. If the algorithm is required to produce K clusters then there will be K initial means and K final means. In completion, K-means algorithm produces K final means which answers why the name of algorithm is K-means.

After termination of K-means clustering, each object in dataset becomes a member of one cluster. This cluster is determined by searching all over the means in order to find the cluster with nearest mean to the object. Shortest distanced mean is considered to be the mean of cluster to which observed object belongs. K-means algorithm tries to group the items in dataset into desired number of clusters. To perform this task it makes some iteration until it converges. After each iteration, calculated means are updated such that they become closer to final means. And finally, the algorithm converges and stops performing iterations.

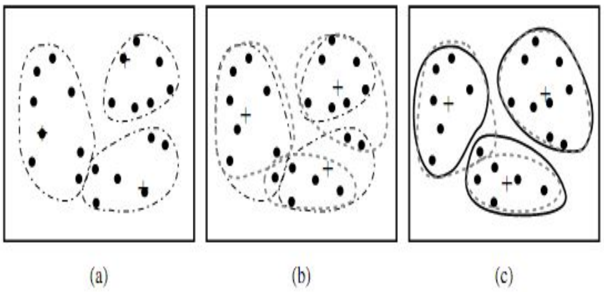

Fig.1 Example of K-means Algorithm [1]

Steps of Algorithm:

```
Input:
D = {d1, d2, dn} //set of n data items.
k // Number of desired clusters
Output: A set of k clusters.
Steps:
1. Arbitrarily choose k data-items from D as initial centroids;
2. Repeat
Assign each item di to the cluster which has the closest centroid;
Calculate new mean for each cluster;
Until convergence criteria is met.
```

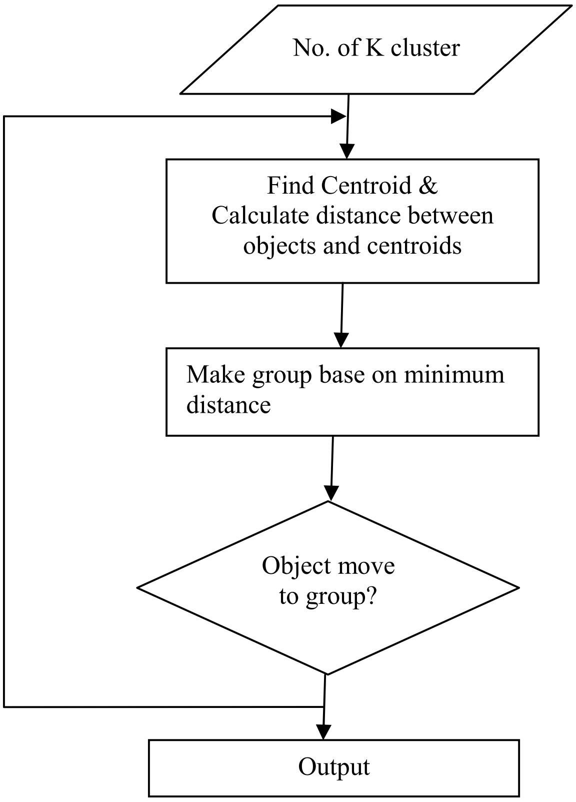
No. of K cluster
Find Centroid & Calculate distance between objects and centroids
Make group base on minimum distance
Object move to group?
Output

Fig.2 Flowchart of K-mean clustering algorithm

Advantages of K-mean clustering

- K-mean clustering is simple and flexible.
- K-mean clustering algorithm is easy to understand and implements.

Disadvantages of K-mean clustering

- In K-mean clustering user need to specify the number of cluster in advance [3].
- K-mean clustering algorithm performance depends on a initial centroids that why the algorithm doesn’t have guarantee for optimal solution [3].

III. Related Work

Paper Title & Approach

1. An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points

D. Napoleon & P. Ganga lakshmi [11] proposed a method for making the K-Means algorithm more valuable and professional; so as to get better clustering with abridged complexity. The best algorithm was found out based on their recital using uniform distribution data points. The accuracy of the algorithm was investigated during different execution of the program on the input data points. From the experimental results, the uniform distribution data points are used easily realize the values and get a superior result.

2. An efficient enhanced k-means clustering algorithm

Fahim et al. [4] proposed k-means algorithm determines spherical shaped cluster, whose center is the magnitude center of points in that cluster, this center moves as new points are added to or detached from it. This proposition makes the center closer to some points and far apart from the other points, the points that become nearer to the center will stay in that cluster, so there is no need to find its distances to other cluster centers. The points far apart from the center may alter the cluster, so only for these points their distances to other cluster centers will be calculated, and assigned to the nearest center.

3. MK-means - Modified K-means clustering algorithm

Hesam et al. [5] proposed The MK-Means clustering uses principal components analysis, to establish a provisional value of count of classes and grant changeable labels for objects. This section is paying interest on how to project objects into space of principal vector and how to discover labels of objects by using a probability of connectivity matrix that for every two objects shows the probability of being in same class. We consider the number of connected objects to Xi from two nearest classes, and also the sum of distances of Xi for connected objects with respect to their class ID.

4. A Novel Density based improved k-means Clustering Algorithm – Dbkmeans

K. Mumtaz et al. [6] proposed in our imitation of the algorithm, we only assumed overlapped clusters of circular or spheroid in nature. So the criteria for splitting or joining a cluster can be determined based on the number of estimated points in a cluster or the estimated density of the cluster.

5. A Survey on K-mean Clustering and Particle Swarm Optimization

P.Vora et al. [7] proposed James MacQueen, the one who proposed the term "k-means"in 1967. But the standard algorithm was firstly introduced by Stuart Lloyd in 1957 as a technique pulse-code modulation. The K-Means clustering algorithm is a partition-based cluster analysis method. According to the algorithm we firstly select k objects as initial cluster centres, then calculate the distance between each cluster centre and each object and allocate it to the nearest cluster, revise the averages of all clusters, repeat this process until the criterion function congregated.

6. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm

Shi Na et al. [8] proposed an improved k-means algorithm to solve shortcomings of standard k-means clustering algorithm, requring a simple data structure to store some information in every iteration,which is to be used in the next interation.The improved method avoids computing the distance of each data object to the cluster centers repeatly, saving the running time.

7. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values

The k-means algorithm is well known for its efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. Zhexu Huang [9] proposed two algorithms which extend the k-means algorithm to categorical domains and domains with mixed numeric and categorical values.

8. Research on Modified k-means Data Cluster Algorithm

Sun Shibao et al. [10] proposed an improved k-means algorithm based on weights. This is a new partitioning clustering algorithm, which can handle the data of numerical attribute, and it also can handle the data of symbol attribute. Meanwhile, this method decreases the impact of remote points and the “noise”, so it improves the efficiency of clustering. However, this method has no improvement on the complexity of time.

IV. Proposed Work

Steps of Algorithm:

Input:
D = {d1, d2,......,dp} // set of p data items
k // Number of needed clusters
Output: A set of k clusters.

Steps:
Phase 1 : Find out the primary centroids of the clusters by using Phase 1.
Phase 2 : Distribute each data point to the suitable clusters by using Phase 2

Phase 1: Find out the primary centroids of the clusters

Input:
D = {d1, d2,......,dp} // set of p data items
k // Number of required clusters
Output: A set of k primary centroids .

```
Steps:
1. Set m = 1 and calculate the distance between all other data-
   points in the      set D;
2. Calculate the adjoining pair of data points from the set D
   and outline a data-point set Am ($1 \leq m \leq k$) which manages
   these two data- points and Delete it from the set D;
3. Find out the data point in D that is nearby to the data point
   set Am, Insert it to Am and Remove it from D; repeat until
   the number of data points in Am reaches $\alpha * (p/k)$;
4. If m<k, then m = m+1, discover one more pair of data
   points from D among which the distance is the shortest,
   form one more data-point set Am and remove them from
   D, Go to step 4;
5. For data-point set Am ($1 \leq m \leq k$) discover the arithmetic
   mean, which will be the primary centroids.
```

Phase 2: Distribute each data point to the suitable clusters

Input:
D = {d1, d2, …. , dp} // set of p data-points.
Ce = {c1, c2, …. , ck} // set of k centroids.
Output: A group of k clusters

```
Steps:
1.  Calculate the distance of all data-point dj (1<= j<= p) to
    all the centroids ci (1<= i<= k) as d(dj,ci);
2.  For all data-point dj, find the nearby centroid ci and assign
    dj to cluster i and Set ClusterId[j] = i; // i:Id of the
    nearby cluster and Nearest_Dist[j] = d(dj,ci);
3.  For all cluster i (1<= i<= k) , recalculate the centroids;
Repeat
4.      For all data-point
    4.1 Calculate its space from the centroid of the current
        nearest cluster;
    If this space is less than or equal to the present nearest
        space, the data-point stays in the cluster;
    Else
    4.1.1     For all centroid ci (1<= i<= k) calculate the
        distance d(dj,ci);
    Endfor;
    4.1.2     Allocate the data-point dj to the cluster with
        the nearby centroid ci. And set ClusterId[j] = i and
        Nearest_Dist[j] = d(dj,ci);
    End for;
5.  For all cluster i (1<= i<= k), repeat to calculate the centroids;
    until the convention criteria is met.
```

Phase 1 describes the method for finding initial centroids of the cluster. Initially, compute the distance between each data point and all other data points in the set of data points. Then find out the closest pair of data points and form a set A1 consisting of these two data points, and delete them from the data point set D. Then determine the data point which is closest to the set A1, add it to set A1 and delete it from D. Repeat this until the number of elements in the set A1 reaches a thresold.

At that point go back to next step and form another data-point set A2. Repeat this till 'q' such sets of data points is obtained. The distance between one vector X = (x1,x2,…xn) and another vector Y = (y1, y2, …, yn) is obtained as,

$d(x,y) = \sqrt{((x1-y1)^2 + (x2-y2)^2 + \ldots + (xn-yn)^2)}$

The initial centroids of the clusters are given as input to the second phase, for assigning data-points to appropriate clusters. The steps involved in this phase are outlined as phase 2.

In phase 2, the first step is to determine the distance between each data-point and the initial centroids of all the clusters. The data-poins are then assigned to the clusters having the closest centroids. For each data-point, the cluster to which it is assigned (ClusterId) and its distance from the centroid of the nearest cluster (Nearest_Dist) are noted. For each cluster, the centroids are recalculated by taking the mean of the values of its data-points. Up to this step, the procedure is almost similar to the original k-means algorithm except that initial centroids are computed systematically.

The next step is an iterative process which makes use of a heuristic method to improve the efficiency. During the iteration, the data-points may get redistributed to different clusters. The method involves keeping track of the distance between each data-point and the centroid of its present nearest cluster. The loop is repeated until no more data-points cross cluster boundaries, which indicates the convergence criterion. The heuristic method described above results in significanct reduction in the number of computations and thus improves the efficiency.

V. Compexity

As we discussed before, the k-means algorithm converges to local minimum. Before the k-means converges, the centroids computed number of times, and all points are assigned to their nearest centroids.i.e. complete redistribution of points according to new centroids, this takes *O*(*pkt*), where p is the number of points, k is the number of clusters and t is the number of iterations.

In the proposed enhaced k-means clustering algorithm, to obtain initial clusters, this process requiers *O*(*pk*). Here, some points remain in its cluster, the others move to another cluster. If the point stays in its cluster this require *O*(1), otherwise require *O*(*k*). If we suppose that half points move from their clusters, this requires *O*(*pk*/2), since the algorithm converges to local minimum, the number of points moved from their clusters decreases in each iteration. So we expect the total cost is $pk\sum_{i=1}^{t} 1/i$. Even for large number of iterations, $pk\sum_{i=1}^{t} 1/i$ is much less than *pqt*. So the cost of using enhanced k-means algorithm approximately is *O*(*pk*), not *O*(*pkt*).

VI. Conclusion

Data mining in recent years with the database and artificial intelligence developed a new technology, its aim the large amount of data from the excavated useful knowledge, to achieve the effective consumption of data resources. As one important function of data mining, clustering analysis either as a separate tool to discover data sources distribution of information, as well as other data mining algorithms as a preprocessing step, the cluster analysis has been into the field of data mining is an important research topic.

The original k-means algorithm is widely used for clustering large sets of data. But it does not always assurance for good results, as the accuracy of the final clusters depend on the selection of initial centroids. Moreover, the computational complication of the original algorithm is very high because it reassigns the data points a number of times during every iteration of the loop.

VII. Future Work

We are scheduling to wished-for new algorithm for finding the initial centroid and doing the experiments on the dataset which we will taken from UCI Machine Learning Repository and check the result. And also doing experiments on same dataset using the standard K-means algorithm and compare both the results and we try to observe the accuracy & time compexity.

References

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000

[2] A. Jain, M. Murty and P. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Vol.31, No. 3, Sep 1999, pp. 264–323.

[3] Garbriela derban and Grigoreta sofia moldovan, "A comparison of clustering techniques in aspect mining", Studia University, Vol LI, Number1, 2006, pp 69-78.

[4] FAHIM A.M., SALEM A.M., TORKEY F.A., RAMADAN M.A." An efficient enhanced k-means clustering algorithm" J Zhejiang Univ SCIENCE A 2006 7(10):1626-1633

[5] Hesam T. Dashti, Tiago Simas, Rita A. Ribeiro, Amir Assadi and Andre Moitinho "MK-means - Modified K-means clustering algorithm" ,*IEEE* ,978-1-4244-8126-2/10/$26.00 ©2010

[6] K. Mumtaz and Dr. K. Duraiswamy," A Novel Density based improved k-means Clustering Algorithm – Dbkmeans" *International Journal on Computer Science and Engineering* ISSN: 0975-3397 213 Vol. 02, No. 02, 2010, 213-218

7] Pritesh Vora and Bhavesh Oza "A Survey on K-mean Clustering and Particle Swarm Optimization", International Journal of Science and Modern Engineering (*IJISME*) ISSN: 2319-6386, Volume-1, Issue-3, February 2013

[8] Shi Na, Liu Xumin, Guan Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm", *Third International Symposium on Intelligent Information Technology and Security Informatics*

[9] Zhexu Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", *Data Mining and Knowledge Discovery* 2, 283–304 (1998)

[10] Sun Shibao, Qin Keyun,"Research on Modified k-means Data Cluster Algorithm", Computer Engineering, vol.33, No.13, pp.200–201, July 2007.

[11] D. Napoleon & P. Ganga lakshmi "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points" IEEE,2010,pp,42-45

[12] Neha Aggarwal, Kirti Aggarwal, Kanika gupta " Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data minig", IJSER, Vol.3,Issue 3,August 2012

[13] M.P.S Bhatia & Deepika Khurana "Analysis of Initial Centers for k-Means Clustering Algorithm" ,IJCA, Volume 71 – No.5, pp 9 – 13, May 2013

Author -Bangoria Bhoomi M. is pursuing M.E. in Computer Engineering from Gujarat Technological University,Gujarat, India.Her research is focused on k-means clustering in Data Mining.