IMPROVED K-MEANS ALGORITHM BASED ON DISSIMILAR VALUES

[1]Kinjal Thakar
[1]Assistant Professor
[1]Department of Information Technology
[1]Silver Oak College of Engineering & Technology, Ahmedabad, India

Abstract : The clustering techniques are the most important part of the data analysis and k-means is the oldest and popular clustering technique used. The paper discusses the traditional K-means algorithm with advantages and disadvantages of it. It also includes researched on enhanced k-means proposed by various authors and it also includes the techniques to improve traditional K-means for better accuracy and efficiency. There are two area of concern for improving K-means; 1) is to select initial centroids and 2) by assigning data points to nearest cluster by using equations for calculating mean and distance between two data points. The time complexity of the proposed K-means technique will be lesser that then the traditional one with increase in accuracy and efficiency. The main purpose of the article is to proposed techniques to enhance the techniques for deriving initial centroids and the assigning of the data points to its nearest clusters. The clustering technique proposed in this paper is enhancing the accuracy and time complexity but it still needs some further improvements and in future it is also viable to include efficient techniques for selecting value for initial clusters(k). Experimental results show that the improved method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the k-means.

IndexTerms **- Data Analysis, Clustering, k-means Algorithm, Clustering Algorithm**

I. 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[3]. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups[9]. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. These challenges led to the emergence of powerful broadly applicable data mining clustering methods surveyed below.

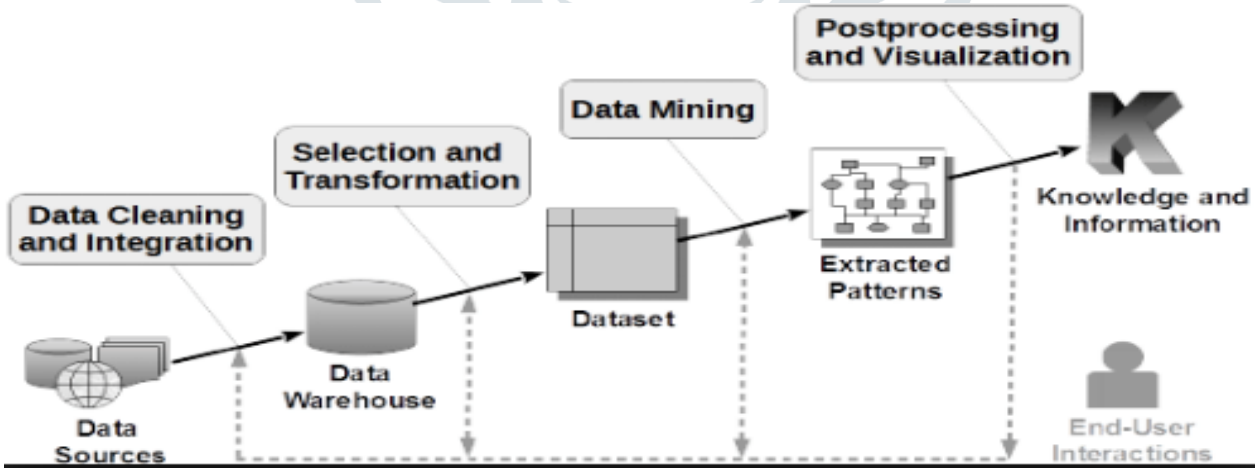

fig 1. data mining process

Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign. For example, a company that sales a variety of products may need to know about the sale of all of their products in order to check that what product is giving

extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving less sale then only the cluster of such products would have to be checked rather than comparing the sales value of all the products.

This is actually to facilitate the mining process. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions.

II. TYPES OF CLUSTERING

Data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at time. Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

2.1 HIERARCHICAL CLUSTERING

A key step in a hierarchical clustering is to select a distance measure. A simple measure is manhattan distance, equal to the sum of absolute distances for each variable. The name comes from the fact that in a two-variable case, the variables can be plotted on a grid that can be compared to city streets, and the distance between two points is the number of blocks a person would walk.

2.2 PARTITIONAL CLUSTERING

Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. This algorithm typically determines all clusters at once. Most applications adopt one of two popular heuristic methods like

2.2.1 K-means algorithm

In k-means case a cluster is represented by its centroid, which is a mean (usually weighted average) of points within a cluster. This works conveniently only with numerical attributes and can be negatively affected by a single outlier. The k-means algorithm is by far the most popular clustering tool used in scientific and industrial applications. The name comes from representing each of k clusters C by the mean (or weighted average) c of its points, the so-called centroid. While this obviously does not work well with a categorical attributes, it has the good geometric and statistical sense for numerical attributes. The sum of discrepancies between a point and its centroid expressed through appropriate distance is used as the objective function. Each point is assigned to the cluster with the closest centroid Number of clusters, K, must be specified. The basic algorithm is very simple.

A. Choose K as the number of clusters.
B. Initialize the codebook vectors of the K clusters (randomly, for instance)
C. For every new sample vector:
    a. Compute the distance between the new vector and every cluster's codebook vector.
    b. Re-compute the closest codebook vector with the new vector, using a learning rate that decreases in time.

Although k-means has the great advantage of being easy to implement, it has some drawbacks. The quality of the final clustering results of the k-means algorithm highly depends on the arbitrary selection of the initial centroids. In the original kmeans algorithm, the initial centroids are chosen randomly and hence we get different clusters for different runs for the same input data [10]. Moreover, the k-means algorithm is computationally very expensive also. The computational time complexity of the k-means algorithm is O(nkl), where n is the total number of data points in the dataset, k is the required number of clusters and l is the number of iterations [2]. So, the computational complexity of the k-means algorithm is rely on the number of data elements, number of clusters and number of iterations.

2.3 DENSITY-BASED CLUSTERING

Density-based algorithms are capable of discovering clusters of arbitrary shapes. Also this provides a natural protection against outliers. These algorithms group objects according to specific density objective functions .Density is usually defined as the number of objects in a particular neighborhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighborhood exceeds some parameter.

2.4 GRID-BASED CLUSTERING

These focus on spatial data i.e the data that model the geometric structure of objects in the space, their relationships, properties and operations. this technique quantize the data set inti a no of cells and then work with objects belonging to these cells. They do not relocate points but ratter build several hierarchical levels of groups of objects. The merging of grids and consequently clusters, does not depend on a distance measure .It is determined by a predefined parameter.

2.5 MODEL-BASED CLUSTERING

Model-Based Clustering methods attempt to optimize the fit between the given data and some mathematical model. Such methods often based on the assumption that the data are generated by mixture of underlying probability distributions. Model-Based Clustering methods follow two major approaches: Statistical Approach or Neural network approach

III. ANALYSIS OF THE PERFORMANCE OF K-MEANS ALGORITHM

3.1 Advantages:

1. K-mean value algorithm is a classic algorithm to resolve cluster problems; this algorithm is relatively simple and fast.

2. For large data collection, this algorithm is relatively flexible and high efficient, because the Complexity is O (ntk). Among which, n is the times of iteration, k is the number of cluster, t is the times of iteration. Usually, k"n and t"n. The algorithm usually ends with local optimum.

3. It provides relatively good result for convex cluster.

4. Because the limitation of the Euclidean distance. It can only process the numerical value, with good geometrical and statistic meaning.

3.2 Disadvantages:

The inherent prosperities of the K-means clustering algorithm to determine its limitations, specific performance is as follows:

1. The K value is most important for K-means clustering algorithm. There is no applicable evidence for the decision of the value of K (number of cluster to generate), and sensitive to initial value, for different initial value, there may be different clusters generated.

2. K-means clustering algorithm has a higher dependence of the initial cluster centers. If the initial cluster center is completely away from the cluster center of the data itself, the number of iterations tends to infinity, but also makes it easier for the final clustering results into local optimization, resulting in incorrect clustering results.

3. K-means clustering algorithm has a strong sensitivity to the noise data objects. If there is a certain amount of noise data in dataset, it will affect the final clustering results, leading to its error.

4. K-means clustering algorithm for the discovery of clusters of arbitrary shape is most difficult.

5. K-means clustering algorithm has man limitation on amount of data. In the iterative process, every time you need to adjust the cluster to which data object belongs and compute cluster center, so in case of large amount of data, the K-means clustering algorithm is not applicable.

IV. IMPROVED K-MEANS CLUSTERING ALGORITHM AT AN ADVANCED LEVEL

Algorithm:

Step1.1: Input Dataset
Step1.2: Check the Each attributes of the Records
Step1.3: Find the mean value for the given Dataset.
Step1.4: Find the distance for each data point from mean value using Equation (Equ).
IF
The Distance between the mean value is minimum then it will be stored in
Then Divide datasets into k cluster points don't needs to move to other clusters.
ESLE
Recalculate distance for each data point from mean value using Equation (Equ) until divide datasets into k cluster

Part2: Assigning data points to nearest centroids
Step2.1: Calculate Distance from each data point to centroids and assign data points to its nearest centroid to form clusters and stored values for each data.
Step2.2: Calculate new centroids for these clusters.
Step2.3: Calculate distance from all centroids to each data point for all data points.
IF
The Distance stored previously is equal to or less then Distance stored in Step2.1
Then Those Data points don't needs to move to other clusters.
ESLE
From the distance calculated assign data point to its nearest centroid by comparing distance from different centroids.
Step2.4: Calculate centroids for these new clusters again.
Until
The convergence criterion met.

V. RESULTS AND DISCUSSION

The traditional K-means clustering is most used technique but it depends on selecting initial centroids and assigning of data points to nearest clusters. There are more advantages than disadvantages of the k-means clustering but it still need some improvements. This paper explains the techniques that improves the techniques for determining initial centroids and assigning data points to its nearest clusters with more accuracy with time complexity of O(n) which is faster than the traditional k-means. The initial value for the K (number of clusters) is still area of concern because it can improve accuracy of the clustering, which will be improved by enhancing the traditional way in future. However, researching on the improvement of K-means clustering algorithms are still not solved completely. And the further attempt and explore will be needed.

REFERENCES

[1] Ghousia Usman, Usman Ahmad and Mudassar Ahmad 2013, Improved K-Means Clustering Algorithm by Getting Initial Cenroids, World Applied Sciences Journal 27 (4): 543-551

[2] Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa2010, Enhancing K-means Clustering Algorithm with Improved Initial Center, International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 121-125

[3] T. Soni Madhulatha 2012, An Overview On Clustering Methods, IOSR Journal of Engineering Vol. 2(4) pp: 719-725

[4] Madhuri A. Dalal, Nareshkumar D. Harale, Umesh L.Kulkarni, 2011, An Iterative Improved k-means Clustering ACEEE Int. J. on Network Security , Vol. 02, No. 03

[5] K. A. Abdul Nazeer, M. P. Sebastian, 2009 Improving the Accuracy and Efficiency of the k-means Clustering Algorithm Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[6] Angélica Urrutia , Hector Valdes , and José Galindo 2013, Comparison of K-Means and Fuzzy C-Means Data Mining Algorithms for Analysis of Management Information: An Open Source Case Computing & Artificial Intelligence, AISC 217, pp. 187–195. Springer International Publishing Switzerland 2013

[7] Anil K. Jain 2010, Data clustering: 50 years beyond K-means, Pattern Recognition Letters 31 (2010) 651–666

[8] Pradeep Rai, Shubha Singh 2010, A Survey of Clustering Techniques International Journal of Computer Applications (0975 – 8887) Volume 7– No.12,

[9] Chunfei Zhang, Zhiyi Fang 2013, An Improved K-means Clustering Algorithm, Journal of Information & Computational Science 10: 1 (2013) 193–199

[10] D T Pham, S S Dimov, and C D Nguyen , 2004, Selection of K in K-means clustering

[11] Shi Na, Liu Xumin, Guan yong, 2010 Research on k-means Clustering Algorithm IEEE DOI 10.1109/IITSI.2010.74