ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

Survey on Various Enhanced K-Means Algorithms

Twinkle Garg[1], Arun Malik[2]

M-Tech Scholar, Dept. Of Computer Science, Lovely Professional University, Jalandhar, India [1]

Assistant Professor, Dept. Of Computer Science, Lovely Professional University, Jalandhar, India [2]

Abstract: Data Mining is defined as a technique used to extract and mine the invisible, meaningful information from mountain of data. Clustering is an important technique that has been introduced in the area of data mining. Clustering is defined as a method used to group similar data into a set of clusters based on some common characteristics. K-means is one of the popular partitional based clustering algorithms in the area of research. The impact factor of k-means is its simplicity, high efficiency and scalability. However, is also comprises of number of limitations: random selection of initial centroids, number of cluster K need to be initialized and influence by outliers. In view of these deficiencies, this paper presents a survey of improvements done to traditional k-means to handle such limitations.

Keywords: Data Mining, Clustering, K-means algorithm, Improved K-means algorithm

1. INTRODUCTION

Data Mining is a technique used to extract and mine the invisible, meaningful information from mountain of data. The term data mining is also relevantly used as Knowledge Discovery in Database, Knowledge engineering. Based on the patterns we look for the Data Mining models and tasks are divided into two main categories Predictive models and Descriptive Models[1].Whereas the Predictive Model is used to predict the feasibility of outcome ,the other Descriptive model is used to describe the important features of dataset. The types of Predictive model are classification, regression ,prediction and time series analysis. The various models included in descriptive model are clustering, summarization, Association rules and sequence discovery. Clustering an unsupervised learning technique established in the area of data mining .

Clustering or cluster analysis can be defined as a data reduction tool used to create subgroups that are more manageable than individual datum. Generally, clustering is defined as a process used for organizing/grouping a large amount of data into meaningful groups or clusters based on some similarity between data. Clusters are the groups that have data similar on basis of common features and dissimilar to data in other clusters. The applications areas where clustering plays an important role are machine learning, image processing, data mining, marketing, text mining. The terms clustering and classifications are always confused with each other, since they are two separate terms. Whereas Clustering is unsupervised learning process because the resulting clusters are not known before the execution which implies the absence of predefined classes in clustering. On the other hand classification is a supervised learning process due to presence of predefined classes. The high quality clustering is to obtain high intracluster similarity and low inter-cluster similarity [2]. There are number of clustering algorithms that are used to cluster the data. These are as shown below:

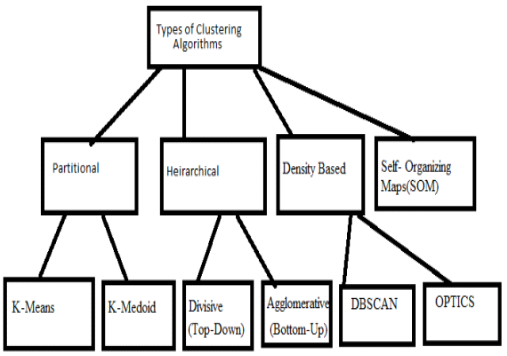

Fig 1: Types of Clustering Algorithms

2. K-MEANS ALGORITHM

K-means is one of the classical partitional based clustering algorithm introduced in the field of data mining. The algorithm was proposed by Mac Queen in the year 1967[5].It was introduced to solve various clustering problems. The algorithm aims to group data into k clusters based on randomly selected initial centroids. The grouping is done by minimizing the Euclidean distances between the data items and its related centroid.K-means algorithm itself is unsupervised and iterative in nature. The clusters generated by k-means are non-heirarchical in nature.

The steps of k-means algorithm are as follows[3]

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move.

This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The complexity of the algorithm is Complexity is

O(n * K * I * d)

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

Where n = number of points, K = number of clusters,
I = number of iterations, d = number of attributes

Advantages of k-means

-It is simple and robust
-If large number of variables exists, then K-Means is computationally faster than hierarchical clustering,if we keep k smalls.
-If the clusters are globular,K-Means produce tighter clusters than hierarchical clustering
-More efficient than k-mediod

Disadvantages

-Difficulty in comparing quality of the clusters produced
-Fixed number of clusters can make it difficult to predict what K should be[3]
-Does not work well with non-globular clusters.
-Different initial partitions can result in different final clusters[8]
-Sensitive to outliers

3. RELATED WORK

Wang Shunye et al[2]Motivated by the problem of random selection of initial centroid and similarity measures ,the researcher presented a new K-means clustering algorithm based on dissimilarity.This improved k-means clustering algorithm basically consists of 3 steps.The first step discussed is the construction of the dissimilarity matrix i.e dm.Secondly ,Huffman tree based on the Huffman algorithm is created according to dissimilarity matrix.The output of Huffman tree gives the initial centroid.Lastly the k-means algorithm is applies to initial centroids to get k cluster as output.Iris,Wine and Balance Scale datasets are selected from UIC machine learning repository to test the proposed algorithm.Compared to traditional k-means the proposed algorithm gives better accuracy rates and results.

Pallavi Purohit and Ritesh Joshi et al[3] proposed an improved approach for original K-means clustering algorithm due to its certain limitations. The main reason for poor performance of K-means algorithm is selection of initial centroids randomly. The proposed algorithm deals with this problem and improves the performance and cluster quality of original kmeans algorithm.The new algorithm selects the initial centroid in a systematic manner rather than randomly selecting. It first find out the closest data points by calculating Euclidian distance between each data point and then these points are deleted from population and forms a new set. This step is repeated on new set by finding data points that are closest to each other. Performance comparison is done using Matlab tool. The proposed algorithm gives more accurate results and also decreases the mean square distance. But the proposed algorithm works better for dense dataset rather than sparse.
Juntao Wang & Xiaolog[4] discuss an improved k-means clustering algorithm to deal with the problem of oulier detection of existing k-means algorithm.The proposed algorithm uses noise data filter to deal with this problem. Density based outlier detection method is applied on the data to be clustered so as to remove the ouliers.The motive of this method is that the outliers may not be engaged in computation of initial cluster centres. In the next step fast global k-means algorithm proposed by Aristidis Likas is applied to the output generated previously.The results between k-means and improved k-means are compared using Iris,Wine,Abolone datasets.The Factors used to test are clustering accuracy and clustering time. The disadvantage of the improved k-means is that while dealing with large data sets ,it will cost more time.

Fahim A.M,Salem A.M,Torkey F.A,Ramadan M.A[5]proposes an efficient enhanced k-means algorithm to overcome problems in existing k-means.Original k-means is famous due to its ease,simplicity,speed of convergence and adaptability to sparse data.Inspite of its large number of advantages ,it suffers from certain disadvantages. These problem are the initialization of centroids,problem to converfe to local minimum i.e updation of centroids till local minimum is not fount & execution of repeated while loops.All these problems are handled by the proposed k-means clustering algorithm.The enhanced algorithm firstly assign datasets to its closest centroid and then compute distance with other centroids.In next step the two distances are compared and if the new distance is smaller than the previous distance then the datapoint is moved to new cluster otherwise if is small then it is assigned to same cluster.This process will save a lot of time and improve the efficiency.This algorithm uses two new functions .The first one is distance() function that is used to compute distance between each datapoint and its nearest cluster head.The second one is distance_new() function used to compute distance between datapoints and other remaining clusters.The experimental results shows that the enhanced k-means algorithm is much fast and efficient than the original k-means .

Md.SohrabMahmud,Md.MostafizerRahman,and Md.NasimAkhtar[6] gave an algorithm to compute better initial centroids based on heuristic method.The newly presented algorithm results in highly accurate clusters with decrease in computational time.In this algorithm author firstly compute the average score of each data points that consists of multiple attributes and weight factor.Merge sort is applied to sort the output that was previously generated.The data points are then divided into k cluster i.e number of desired cluster.Finally the nearest possible data point of the mean is taken as initial centroid.Experimental results shows that the algorithm reduces the number of iterations to assign data into a cluster.But the algorithm still deals with the problem of assigning number of desired cluster as input.
Shuhua Ren and Alin Fan [7] elobrates k-means clustering algorithm based on coefficient of variation.The coefficient of variation is defined as ratio of standard deviation to the mean value.Existing k-means algorithm uses Euclidean distance as the similarity metric which gives inacurate results due to the effect of useless data. To overcome with this problem ,proposed algorithm uses coefficient of weight factor to elicit the effect of ouliers.Weight values

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

are assigned to all the features in clustering to remove irrelevant,noisy data so as to increase cluster quality.The results are evaluated using popular data sets i.e. Iris,Wine and Balance scale.The results prove that the modified algorithm presents more clustering accuracy and the number of iterations required for clustering are less than original k-means.The problem faced by proposed algorithm is that the number of clusters required as output are needed to be initially defined.

Navjot Kaur, Navneet Kaur[8]enhaced the traditional k-means by introducing Ranking method.Author introduces Ranking Method to overcome the deficiency of more execution time taken by traditional k-means.The Ranking Method is a way to find the occurrence of similar data and to improve search effectiveness.The tool used to implemented the improved algorithm is Visual Studio 2008 using C#.The advantages of k-means are also analyzed in this paper.The author finds k-means as fast ,robust and easy understandable algorithm.He also discuss that the clusters are non-heirarchical in nature and are not overlapping in nature.The process used in the algorithm takes student marks as data set and then initial centroid is selected.Euclidean distance is then calculated from centroid for each data object.Then the threshold value is set for each data set.Ranking Method is applied next and finally the clusters are created based on minimum distance between the data point and the centroid.The future scope of this paper is use of Query Redirection can be used to cluster huge amount of data from various databases.

Raju G, Binu Thomas, Sonam Tobgay and Th. Shanta Kumar[9] gave a comparative analysis between k-means clustering algorithm and fuzzy clustering algorithm.In this paper the researcher also discuss the advantages and limitations of fuzzy c-means algorithm.K-means is a partional based clustering algorithm whereas Fuzzy c-means is non partional based clustering algorithm.Fuzzy c-means mainly works in two process.In the first process cluster centers are calculated and in second the data points are assigned to calculated cluster center with the help of Euclidean distance.This process is almost similar to conventional k-means with a little difference.In fuzzy c-means algorithm membership value ranging from 0 to 1 is assigned to data item in cluster.0 membership indicates that the data point is not a member of cluster whereas 1 indicates the degree to which data point represents a cluster.The problem faced by fuzzy c-means algorithm is that the sum of membership value of data points in each cluster is restricted to 1.Algorithm also face problem in dealing with outliers.On the other hand comparison with k-means shows that the fuzzy algorithm is efficient in obtaining hidden patterns and information from natural data with outlier points

4. CONCLUSION

Clustering play a crucial role in many applications.The commonly used efficient clustering algorithm is k-means clustering.K-means clustering is an important topic of research now a days in data mining. This paper have presented a survey of most recent research work done in this area.However k-means is still at the stage of exploration and development.The survey concludes that many improvements are basically required on k-means to improve problem of cluster initialization,cluster quality and efficiency of algorithm.

ACKNOWLEDGMENT

The author is very thankful to Asst. Professor Arun Malik for his guidance for paper and other faculties for their suggestions in paper.

REFERENCES

[1] Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES:SURVEY PAPER"IJRET eISSN: 2319-1163 | pISSN: 2321-7308

[2] Wang Shunye "An Improved K-means Clustering Algorithm Based on Dissimilarity" 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)Dec 20-22, 2013, Shenyang, China IEEE

[3] Pallavi Purohit "A new Efficient Approach towards k-means Clustering Algorithm" ,International journal of Computer Applications,Vol 65-no 11,march 2013

[4] Juntao Wang & Xiaolong Su "An improved K-Means clustering algorithm" 2011 IEEE

[5] FAHIM , SALEM A.M, TORKEY F.A, RAMADAN M.A "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University SCIENCE A ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)

[6] Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md.Nasim Akhtar "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average" 2012 7th International Conference on Electrical and Computer Engineering 20-22 December, 2012, Dhaka, Bangladesh, 2012 IEEE

[7] Shuhua Ren &Alin Fan "K-means Clustering Algorithm Based On Coefficient Of Variation" 2011 4th International Congress on Image and Signal Processing2011 IEEE

[8] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur "EFFICIENT K-MEANSCLUSTERING ALGORITHM USING RANKING METHOD IN DATA MINING" ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012

[9] Raju G, Binu Thomas, Sonam Tobgay and Th. Shanta Kumar"Fuzzy Clustering Methods in Data Mining:A comparative Case Analysis" 2008 International Conference on advanced computer theory and engineering,2008 IEEE

BIOGRAPHIES

Twinkle Garg, M-tech(pursuing),Department of Computer Science &Technology, Lovely Professional University, Jalandhar, India. Area of research Data Mining

Arun Malik, Asst Professor, Department of Computer Science &Technology, Lovely Professional University, Jalandhar ,India