A Survey Paper on Improvisation of K-Means Clustering Algorithm with Implementation on E-Commerce Data

Shekh Shabazhusen A[1] Prof.Ketan Patel[2]
[1,2]G.M.F.E, Himmatnagar

Abstract— Clustering is a technique for primary data analysis and k-means is clustering algorithms which are very useful of all the other algorithms. In data mining clustering methods are frequently used, because its performance in clustering large data sets. The result of the k-means clustering algorithm depends upon the correctness of the initial centroids, because they 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 research 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. Our experimental results show the proposed algorithm more accurate with less computational time comparing to original k-means clustering algorithm.

Key words: Text Summarization, Key phrases Extraction, Text mining, Data Mining and Text compression

I. INTRODUCTION

Clustering technique is a way that classifies the raw data reasonably and it searches the hidden patterns from datasets which are existing.

In this process of grouping data objects into disjointed clusters so that the data in the same cluster are similar, yet data belonging to different cluster differ.

The demand for organizing the sharp increasing data and learning valuable information from data, which makes clustering techniques are widely applied in many application areas such as artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition, psychology, statistics and so on.

The algorithm is a numerical, unsupervised, non-deterministic, iterative method. It is simple as well very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. But it is very suitable for producing globular clusters. Several attempts were made by researchers to improve efficiency of the k-means algorithms.

The algorithm of this paper proposed is superior to the standard k-means method on running time and accuracy, thus enhancing the speed of clustering and improving the time complexity of algorithm. By comparing the experimental results of the standard k-means and the improved k-means, we can see that the improved method can effectively shorten the running time. 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 clustering does not depend on predefined classes and training examples while classifying the data objects. Cluster analysis seeks to partition a given dataset 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 a crucial area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics, etc.

II. THE PROPOSED SYSTEM ARCHITECTURE

The following diagram figure.1 represents the proposed system:

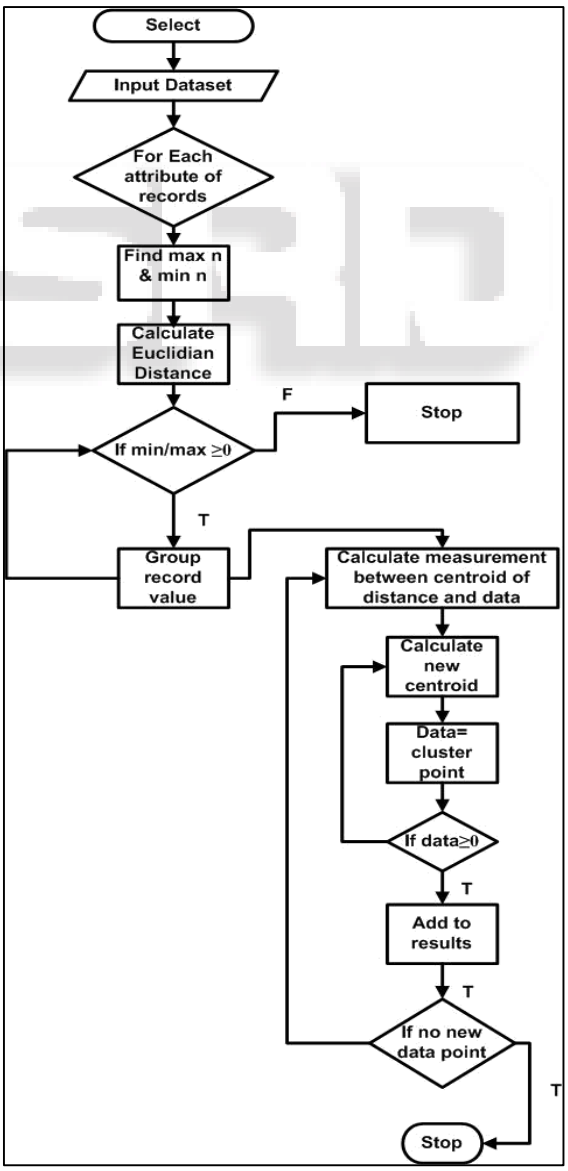

Fig. 1: Proposed Flowchart

III. PROPOSED ALGORITHM

Step 1:- Select
Step 2:- Input Dataset for each attributes records
Step 3:- Find max n & min n
Step 4:- Calculate Euclidean Distance
Step 5:- If min/max ≥ 0 then record group value
Step 6:- Return to step 5
Step 7:- If min/max >≠0 then decide dataset into k clusters
Step 8:- Calculate centroid
Step 9:- Calculate measurement between centroid of distance and data
Step 10:- Calculate new centroid
Step 11:- Where data=clusters points
Step 12:- If data ≥ 0 then add to results
Step 13:- Otherwise go to step 10
Step 14:- If no new data point available then
Step 15:- Stop
Step 16:- Otherwise return to step 9

A. Proposed Workflow

Data reduction is the most important problem to work with huge datasets. To reduce the size ofdataset to value it, we use k-nn classification. Computing Euclidian distance to measure dissimilarity between two distances. There is a drawback of k means where the complexity in search of n-n parameters.

$$\text{Accuracy}=\frac{\#\ of correctly classified examples}{(no of examples in n.k)\times 1000}$$

Where n.k is validation set

Here Euclidian distance is used because k-means is a vector quantization method. It amounts torepeatedly assigning points to the closet centroid so, Euclidian distance from data to a centroid in measured.

$$\text{Euclidian Distance} = \sqrt{\sum_{k=1}^{n}(\text{pk} - \text{qk})\text{^}2}$$

n = no of dimensions (attributes) and p_k andq_k are respectively the k[th]attributes or data object p and q.

Large Dataset = Bank Dataset
Smallest Dataset = Mashroom Dataset

We have observed that provide algorithm have taken multiple factors in consideration and so, it takes time to calculate more nearby centroid and the results are not accurate. So we have tried to reach to more accurate distance between a point and a centroid and better clustering method.

Thus we are importing it by;
Focusing on more closer centroid searching and providing better measurement formula.

IV. ANALYSIS OF THE PERFORMANCE OF K-MEANS ALGORITHM

A. Advantages:

1) To resolve cluster problemsK-mean value algorithm is a classic algorithm and this algorithm isrelativelyfastand simple.
2) The k means is relatively flexible and high efficient big dataset, because theComplexity 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 localoptimum.
3) The result is relatively good for convex cluster.
4) The limitation of the Euclidean distance may process the numerical value,with good geometrical and statistic meaning.

B. Disadvantages:

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

1) The K value is most important for this algorithm. There is no applicableevidence for the decision of the value of K (number of cluster to generate), and sensitive toinitial value, for different initial value, there may be different clusters generated.
2) K-means clustering algorithm has a higher dependence of the initial cluster centres. If theinitial cluster centre is completely away from the cluster centre of the data itself, the numberof iterations tends to infinity, but also makes it easier for the final clustering results intolocal optimization, resulting in incorrect clustering results.
3) K-means clustering algorithm has a strong sensitivity to the noise data objects. If there isa certain amount of noise data in dataset, it will cause the final clustering results and leading to its error.
4) For the discovery of clusters of arbitrary shape is most difficult in K-means clustering algorithm.
5) The main limitation ofK-means clustering algorithm is amount of data. In the iterative process, every time you need to adjust the cluster to which data object belongs and compute clustercentre, so in case of large amount of data, the K-means clustering algorithm is not applicable.

V. THE RESEARCH POINT OF K-MEANS CLUSTERING ALGORITHM

The research on K-means clustering algorithm is mainly from the following two aspects:

First, about the determination of K value. Through the above analysis, the K value of the initial cluster centres to determine the far-reaching impact throughout the clustering process and the final clustering results, while the K value in practical applications is very difficult to direct or one-time determination. Especially, If the pending amount of data tends to infinity, the K value of the algorithm to determine will be very difficult. At present, there are twok mean clustering algorithms to determine the K value is relatively effective which is the cost functionbased on distance and propagation clustering algorithm based on nearest neighbours. The formerfind the minimum through using the cost function. Thus obtain the corresponding K value. Thelatter using nearest neighbour clustering algorithm to calculate the appropriate number of clustercentre, the number of cluster centre provides for the maximum K value of the K-means clustering algorithm to get the optimal value of K.Second, about the choice of initial cluster centres. This algorithm using theiterative method to solve the problem and except the first step then clustering results of each stepare improved to some extent; otherwise terminate the process of iteration. This method takes the cluster squares error and the criterion function value change or notas the iterative termination conditions. But the clustering results obtained from this criterionfunction easily fall into local minimum solution and

the result is the clustering results of search aremoving toward the direction of diminishing the criterion function value .

In this paper, the improvement of K-means algorithm is mainly reflected in the following twoaspects:

- Optimize the initial cluster centres and then find a set of data that reflect the characteristics of datadistribution as the initial cluster centres, to support the division of the data to the greatestextent.
- Optimize the calculation of cluster centres and data points to the cluster centre distance then make it more match with the goal of clustering.

A. *The Process of K-Means Algorithm*

Kmeansis a clustering algorithm in which it is widely used for clustering large dataset. In 1967,MacQueen firstly proposed the k-means algorithm; it was one ofthe most simple, non-supervised learning algorithms, which wasapplied to solve the problem of the well-known cluster. It is apartitioning clustering algorithm, this method is to classify thegiven date objects into k different clusters through the iterative, converging to a local minimum. So the results of generatedclusters are compact and independent.The algorithm consists of two separate phases. The first phaseselects k centres randomly, where the value k is fixed in advance.

In next phase, we take each data object to the nearest centre.Euclidean distance is generally considered to determine thedistance between each data object and the cluster centres. The first step iscompleted and an early grouping is donewhenall the data objects are included in some clusters. Calculating theaverage again of the early formed clusters. This iterative processcontinues repeatedly until the criterion function becomes theminimum.

The process of k-means algorithm as follow:

Input:

Number of desired clusters, k, and a database D= {d1, d2,…dn} containing n data objects.

Output:

A set of k clusters

Steps:

1) Randomly select k data objects from dataset D as initial cluster centres.

2) Repeat;

3) Calculate the distance between each data object di (1 <= i<=n) and all k cluster centres(1<=j<=k) and assign data object di to the nearest cluster.

4) For each cluster j (1<=j<=k), recalculate the clustercentre.

5) Until no changing in the centre of clusters.

B. *The shortcomings of k-means algorithm*

The above analysis of k mean algorithms, the distance can calculate from each data object to every cluster centre in each iteration in this algorithm. However, by experiments we see that it is not necessary for us to calculate that distance each time. Assuming that cluster C formed after the first iterations, the data object x is assigned to cluster C, but in a few iterations, the data object x is still assigned to the cluster C. In this process, after several iterations, we calculate the distance from data object x to each cluster centre and find that the distance to the cluster C is the smallest. So in the course of several iterations, k-means algorithm is to calculate the distance between data object x to the other cluster centre, which takes up a long execution time thus affecting the efficiency of clustering.

VI. RESULTS

This algorithm is also applied to the clustering of real datasets. In two experiments, time taken for each experiment is computed. The same data set is given as input to the standard kmeansalgorithm and the improved algorithm. Experimentscompare improved k-means algorithm with the standard k-meansalgorithm in terms of the total execution time of clusters and theiraccuracy. Experimental operating system is Window XP,program language is VC++ 6.0.In this paper uses iris, glass, letter [4] as the test datasets and also this papergives a brief description of the datasets used in experimentevaluation. Table 1 shows some characteristics of the datasets of k mean algorithm.

Dataset	Number of attributes	Number of records
Iris	4	150
Glass	9	214
letter	16	20000

Table 1: Charactristics of the Datasets

In experiment 1, datasets of Iiris and glass are selectedbecause they are fit to clustering analysis and their clusteringresults are reliable. The number of cluster k sets 3. Clusteringresults for the standard k-means algorithm and the improved kmeansalgorithm proposed in this paper are listed in Table Ⅱ .

Dataset	k-means Running time (s)	Improved k-means Running time(s)	k-means Accuracy %	Improved k-means Accuracy %
Iris	0.0586	0.0559	84.3	91.6
glass	0.0814	0.0778	78.9	89.3

Table 2: Clustering Results for IRIS, Class on the Standard K-Means and the Improved K-Means

The improved k-means algorithm can produce the final cluster results in shorter time than the standard k-means in the results of experiment 1. At the same time the improved k-means can enhance the accuracy of algorithm.

VII. CONCLUSION

This is a typical clustering algorithm and it is widely used for clustering big dataset. This paper elaborates k-mean s algorithm and analyses the results of the standard k-means clustering algorithm. Because the computational complexity of the standard k-means algorithm is objectionably high owing to the need to reassign the data points a number of times during each iteration, which makes the efficiency of standard clustering is low. This paper presents simple and

efficient way for assigning data points to clusters. The proposed method in this paper ensures the entire process of clustering in O(nk) time without sacrificing the accuracy of clusters. Experimental results of the improved algorithm can improve the execution time of k-means algorithm. So the proposed k-means method is feasible as well as simple and fast.

REFERENCES

[1] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004.
[2] Sun Jigui, Liu Jie, Zhao Lianyu, "Clustering algorithms Research",Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.
[3] Sun Shibao, Qin Keyun,"Research on Modified k-means Data Cluster Algorithm"I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Computer Engineering, vol.33, No.13, pp.200–201,July 2007.
[4] Feil, B., Abonyi, J.: Introduction to Fuzzy Data Mining Methods. In: Galindo, J. (ed.) Handbook of Research on Fuzzy Information Processing in Databases, vol. I, pp. 55–95. Information Science Reference, Hershey (2008)
[5] Reddy, G.S., Srinivasu, R., ChanderRao, M.P., Reddy Rikkula, S.: Data warehousing, Data Mining, OLAP and OLTP technologies are essential elements to support decisionmaking process in industries. (IJCSE) Int. Journal on Computer Science and Engineering 2(9), 2865–2873 (2010)
[6] Héctor, V.: Implementación de los Algoritmos de Minería de Datos K-Means y Fuzzy C-Means para el Análisis de Información de Gestión: Un Caso Open Source. Tesis de licenciatura en Ciencias de la Ingeniería, Facultad de Ingeniería Universidad Católicadel Maule, Talca Chile (2012)
[7] Mc Queen J, "Some methods for classification and analysis ofmultivariate observations," *Proc. 5th Berkeley Symp. Math.Statist.Prob.*, (1): 281–297, 1967.
[8] A. Bhattacharya and R. K. De, "Divisive Correlation ClusteringAlgorithm (DCCA) for grouping of genes: detecting varying patternsin expression profiles," *bioinformatics*, Vol. 24, pp. 1359-1366, 2008.
[9] Margaret H Dunham, Data Mining-Introductory and Advanced Concepts, Pearson Education, 2006.
[10] Elmasri, Navathe, Somayajulu, Gupta, Fundamentals of Database Systems, Pearson Education, First edition, 2006.

